SERVER-SIDE ANALYTICS

A BDTF White Paper with accompanying recommendations

Ad Hoc Task Force on Big Data

November 3, 2017

All studies, findings and recommendations in these deliverables have been submitted to the Science Committee and to officials at NASA HQ. The opinions expressed in these materials do not reflect NASA's concurrence, approval, or indicate steps to implementation.

A BDTF White Paper

Server-side Analytics

I. Executive Summary

NASA's archives of science data are growing larger and incorporating more features. In many cases efforts needed by users of these data are being hampered by the inefficiencies associated with moving large data sets from the archives across the internet into the storage local to the user. A new paradigm is beginning to appear which will permit operations on the data prior to transmission. These operations reduce the amount of data that will be transmitted. There are several terms being used to describe the new paradigm: data analytics, climate analytics, server-side analytics, etc., which all have common features – namely bring data analysis code to the data. This paper will use "server-side analytics" (SSA) as being synonymous with the other terms.

The rationale for moving to server-side analytics architectures include:

- Reading data over the internet is slower than reading data from a local file.
- Some analysis tasks require a lot of data: *"Compute the mean annual cycle from a 100-year climate model run."*
- Analysis expression is evaluated at the server, where the data reside on disk.
- Only the analysis result is delivered to the client.
- Server-side analysis saves a lot of time if: (size of result) << (size of data operated on)

This white paper lays out the case for adopting SSA implementations where needed and provides the following recommendations to SMD:

- Publish an RFI via NSPIRES to ascertain where there are centers of SSA development that could be useful for NASA-sponsored research.
- Convene workshops for the NASA science community to show off the SSA services that are under development and get feedback that could point to where new developments are needed,
- Solicit via ROSES for demonstrations or prototypes of SSAs that have the most promise of solving "log jam" situations in priority areas.
- For those projects that demonstrate potential at significantly improving the delivery of tailored processed data to the researcher, SMD should take steps to make those projects operational.

In all events, members of organizations who provide SMD data archive services need to participate. It is most important that representatives, ideas, and approaches from the broader data user communities need to participate and in certain cases lead in the implementations.

Sound science leadership within SMD will be needed to ensure that new SSA architectures are implemented where most needed and other proposals for lower priority implementations are reserved. The use of community peer groups and advisory working groups will be valuable in providing priority rankings for new concepts and proposals.

II. Introduction

NASA-sponsored investigators have a long tradition of serving data derived from their instruments to colleagues in their respective fields. In the decades of the 1970s to the 1980s, this data service involved having the investigator receiving a data request by phone or mail and making a digital data tape employing the group's miniVax or PDP-9. In the 1990s and into the 21st century, with the adoption of the internet and the World Wide Web, this data request and transmission model became more efficient as requesters employed web-based forms and data were transmitted via ftp protocols. Data archives were created as focused centers for holding instrument data and serving the research community at large. This merely codified the old model of data requests and delivery that was developed 40 years ago.

III. Statement of the Problem

The standard method for analyzing Earth and space science data (See Figure 1) is that data residing in one or many data stores must be parsed out and shipped over the internet to the researcher who, in turn, locally stores the data for subsequent analyses. The analysis tasks are varied and include visualization, parameterization, and comparison with or assimilation into physics models.

In many cases this process is inefficient and unwieldy as the data sets become larger and demands on the analysis tasks become more sophisticated and complex. As data storage volumes have reached petabyte size, wide-area networks have not kept pace, so that moving data sets or large subsets from an archive to a user's local device is rapidly becoming impossible.

For about a decade, several groups have explored a new paradigm to this model. The names applied to the paradigm include "data analytics", "climate analytics", "bringing code to data" and "server-side analytics" – all variations of a similar theme. The general idea (See Figure 2) is that there will need to be in close network proximity to certain "big" data stores a tailored processing capability appropriate to the type and use of the data served. The user of the server-side analytics capability will access and retrieve tailored data via the processor that will operate on the data with numerical procedures controlled by the user. Results of the analytics processes will then be relayed to the user. In practice, these results will occupy a much smaller volume, easier to transport to and store locally by the user and easier for the user to interoperate with data sets from other remote data stores. A major component of server-side analytics could be to provide sets of tailored results to end users in order to eliminate the repetitive preconditioning that is both often required with these data sets and which drives much of the throughput challenges.

The main need for SSA is that the linear rate of growth in wide area network capacity is not keeping pace with exponential growth of data storage capacity. Users are increasingly interested in interrogating larger subsets of data archives (e.g., in order to conduct needle-in-haystack big-data analytics with machine learning and other automated procedures), and data archives are reaching beyond petabyte volumes, it is no longer feasible for users to download the data they need over wide area networks. Nor is it realistic to assume that the users have access to the necessary local data stores needed for temporary caching of the research data. They need hundreds of terabytes or even petabytes but often they can only get tens of TB or less.

As an example, one NASA-funded university research group recently acquired a petabyte data storage system. The group intends to do some very complex analyses of data that are residing at NCAR, NASA Ames and the Texas Advanced Computing Center. The new local data storage is providing a great opportunity to bring together massive data sets that are stored at disparate locations. Despite using of sophisticated data transfer optimizations tuned in coordination with the transmitting data centers, the current estimate is that just to download these requested data sets from the 3 remote locations will take upwards of 4 months!

By having SSA in place where the large data sets reside, the operations that would be done on the user's home system instead could be done on the analytics servers available in network proximity to the data archive site. Only the results of analytics processing would be transported back to the user, in digital, tabular or graphical form. Such reductions could result in many orders of magnitude smaller data volumes transferred over the wide area network[1]. One could imagine even more sophisticated analyses – energy budgets, empirical orthogonal functions, probability distributions – being conducted remotely by running automated scripts on the server side, which would result in even larger reductions in data volume to be transported.

Another reason for adopting SSA architectures is that under the current paradigm of having each user download their own copy of a given data set and devise ways of operating on it has great potential for duplication of effort and error. Modern processing streams are typically composed of many steps in which various transformations are applied to the data. It is quite likely that many of the steps are common to many users' intended analyses. As an example, in climate research users typically operate on anomalies, which are the results of subtracting the mean value from the raw field. Thus the calculation of a time mean (which may be time dependent, e.g., if there is a known cycle like the annual cycle of insolation) is repeated by many users. Having each user "invent" the method for analyzing the data means that many users are likely creating code to do the same set of operations. Furthermore, as more users develop their own ways of processing data, the likelihood increases that at least one of them introduces an error. Using the climate analysis example, there are many ways to compute a climatological mean (normal), and some may be more appropriate than others, depending on the intended subsequent processing of anomalies. The promise of SSA is that the processing stream can be developed once, unit tested, and peer-reviewed for accuracy and efficiency, after which it can be made available to the entire user community. By making the SSA library a community code, it also would be possible to capture the methods developed by users and give the code to all other users.

[1] For example, a typical calculation with the data mentioned above involves computing the climatological average, ensemble average and zonal average of a set of, say, 100 different statistics (atmosphere, ocean, land and cryosphere variables; maxes, mins, means, std. deviations, etc.), across ~30-50 years of data, with ensembles of 20 or more members. The data that would be "touched" in such a calculation for a data set with, say 25-km grid spacing (reanalysis data or climate runs) or 5-km grid spacing (satellite data or weather runs) would amount to several 10s of TB. The result would be reduced by 50 * 20 * 1000 = O(10**6) for each statistic, i.e., 10s of MB would be returned to the user, which is a very manageable data volume to move over the wide area network.

A third aspect of the SSA requirement is that interoperability is becoming more and more critically important. If a user wants to do a calculation with data from different archives that are in incompatible formats, a lot of work is needed. Rather than have each user re-discover the way to interoperate on diverse data sets, an SSA implementation could provide scripts that transform data in one format to another to enable interoperability.

This study presents a new architectural approach that may be applicable to several of the four main science domains sponsored by NASA[2]. Investment and development will be needed. These could be provided via a new set of program initiatives issued from SMD.

The analyzing and assimilating large observational data sets will become more efficient leading to more rapid scientific discoveries. A measure of success in adopting the server-side architecture might be that the performance metrics for a data service will incorporate the delivery of meaningful data elements rather than simply volume of bytes transferred.

IV. Discussion and Examples of approaches

The analytics paradigm is quite common and ubiquitously employed – in the apps resident on smart phones, tablets and the like. Here the user hits some keys and icons on the app which encodes and relays the requested action across the internet to the app's home processor. The processor accesses required data which can be local or remote to the processor, prepares its response and transmits back to the app on the user's device – directions to the restaurant, weather in San Francisco, etc.

A very well-known example of a server-side analytics process is Google Earth. The Google Earth application or web site (https://www.google.com/earth/) brings close-up views of the Earth's surface interactively. Google has prepared (and continually updates) a whole-Earth database of images acquired from multiple sources including the NASA-related Earth observatories. The user does not directly go to this database to select and pull the data to their computer. Rather the user's request goes to a special analytics service that (1) requests the image data records from the database and (2) transmits only those records along with coded instructions to the computer's application for rendering to the user. Google has further extended this capability with its "Timelapse" feature (see: https://earthengine.google.com/timelapse/). Similar to Google Earth, Timelapse brings to the user a video rendering of the 32+ years of LandSat images at a point on the Earth's surface that is specified by the user.

New computing tools are arriving and being employed which will facilitate development of SSA implementations. Computational notebooks hosted on the web utilize server-side data and computational resources. For example, Jupyter Hubs (see Figure 3) can enable remote users to run programs, algorithms, and other scripts in common scientific languages like Python, R, and Julia via the user's web browser. (see: http://jupyterhub.readthedocs.io). The programs, workflows, graphs, and annotations are compiled into notebooks, which capture "*…the whole computation process: developing, documenting, and executing code, as well as communicating the results.*" Jupyter Hubs are currently operated by the Department of Energy and other High-

[2] Astrophysics, Earth Science, Heliophysics and Planetary Science. See: https://science.nasa.gov

Performance Computation facilities, and Jupyter Notebooks were used to release the discovery of the recent detection of gravitational waves by LIGO (https://losc.ligo.org/tutorials/). The LSST consortium recently announced that they plan to use JupyterLab for their 'science platform[3].

The NASA research community has explored with some prototyped analytics processing and, in some cases, gone on and implemented specific SSA applications. The APPENDIX has discussions on the following projects: GrADS, MERRA/AS, OceanWorks, Western States Water Mission, SWOT and NISAR science data systems, researcher's tools for accessing and analyzing SDO data, SciServer, the NOAO DataLab and GAVIP.

The implementation of SSA architectures will be of great benefit to many NASA areas of research – but not all. In many of the Earth sciences and solar astronomy research areas initial implementations are showing great promise. Certainly, an accelerated effort will be needed as shown by the examples cited above. NASA's astronomy and astrophysics fields are venturing into the large-area survey mode. In the planetary sciences, the search for both new and previously found Near-Earth Objects most likely could be a fertile ground for implementation of SSA architectures. But in many fields, such as high-energy astrophysics, SSA architectures will not be appropriate since the traditional request-and-retrieve methods are more than adequate for the typical lower-volume data sets.

V. Recommended approach

SMD needs to start exploring implementations for server-side analytics services. Probably the most expeditious and efficient way to initiate this is to publish an RFI via NSPIRES to ascertain where there are centers of SSA development that could be useful for NASA-sponsored research. Also, the SMD should convene workshops for the NASA science community to show off the SSA services that are under development and get feedback that could on where new developments are needed. Based on the results of these two community events, SMD can solicit via ROSES for demonstrations or prototypes of SSAs that have the most promise of solving "log jam" situations in priority areas. For those projects that demonstrate potential at significantly improving the delivery of tailored processed data to the researcher, SMD should take steps to make those projects operational. In all events, members of organizations who provide SMD data archive services need to participate. It is most important that representatives, ideas, and approaches from the broader data user communities need to participate and in certain cases lead in the implementations.

Sound science leadership within SMD will be needed to ensure that new SSA architectures are implemented where most needed and other proposals for lower priority implementations reserved. The use of community peer groups and advisory working groups will be valuable in providing priority rankings for new concepts and proposals.

VI. Conclusion

This white paper lays out the case for adopting server-side analytics implementations. It provides rationale, a generic concept, examples of early prototyping and adaptations and a set of recommendations for SMD to consider.

[3] The LLST consortium has produced a design document for its Jupyter implementation: https://ldm-542.lsst.io

SSA implementations are desperately needed in certain applications, e.g. climatology and solar astronomy. In some cases, a SSA implementation will need to be built into the science operations architecture of a high-data volume mission. There are several fields within the NASA science domain where the SSA implementations do not make sense. Careful attention to community needs and priorities will be needed to focus resources where needed the most.

VII. Acknowledgements

The BDTF appreciates contributions and reviews of drafts of this document by:

Gavin Schmidt, GISS
Alex Szalay, JHU
Arfon Smith, STScI
John Schnase, GSFC
Dan Crichton, JPL
Thomas Huang, JPL
Hook Hua, JPL

VIII. List of Acronyms

AIA	Atmospheric Imaging Assembly
AIST	Advanced Information Systems Technology
API	Application Programming Interface
BDTF	Big Data Task Force (formally Ad-Hoc Task Force on Big Data)
COAPS	Center for Ocean-Atmospheric Prediction Studies
DAAC	Distributed Active Archive Center
DOMS	Distributed Oceanographic Match Service
EDGE	Extensible Data Gateway Environment
EOSDIS	Earth Observing System Data and Information System
GAVIP	Gaia Added Value Interface Platform
GrADS	Grid Analysis and Display System
HEK	Heliophysics Events Knowledgebase
HMI	Helioseismic and Magnetic Imager
JSOC	Joint Science Operations Center
JWST	James Webb Space Telescope
LIGO	Laser Interferometer Gravitational-Wave Observatory
LSST	Large Synoptic Survey Telescope
MERRA	Modern-Era Retrospective analysis for Research and Applications
MUDROD	Mining and Utilizing Data Relevancy from Oceanographic Dataset
NCAR	National Center for Atmospheric Research
NISAR	NASA-ISRO Synthetic Aperture Radar
NOAO	National Optical Astronomy Observatory
NSPIRES	NASA Solicitation and Proposal Integrated Review and Evaluation System
RFI	Request for Information
Pan-STARRS	Panoramic Survey Telescope and Rapid Response System

RFI	Request for Information
ROSES	Research Opportunities in Space and Earth Science
SAR	Synthetic Aperture Radar
SDAP	Science Data Analytics Platform
SDO	Solar Dynamics Observatory
SDS	Science Data System
SDSS	Sloane Digital Sky Survey
SMD	Science Mission Directorate
SSA	Server-Side Analtics (synonymous with Data Analytics, Climate Analytics, etc.)
SWOT	Surface Water Ocean Tomography
TB	Tera Bytes

IX Figures

1. Traditional Data Request and Delivery Model
2. Server-Side Analytics Model
3. Jupyter Hub
4. OceanWorks Architecture
5. WaterTrek Data Analytics Architecture
6. SWOT and NISAR Data Volumes

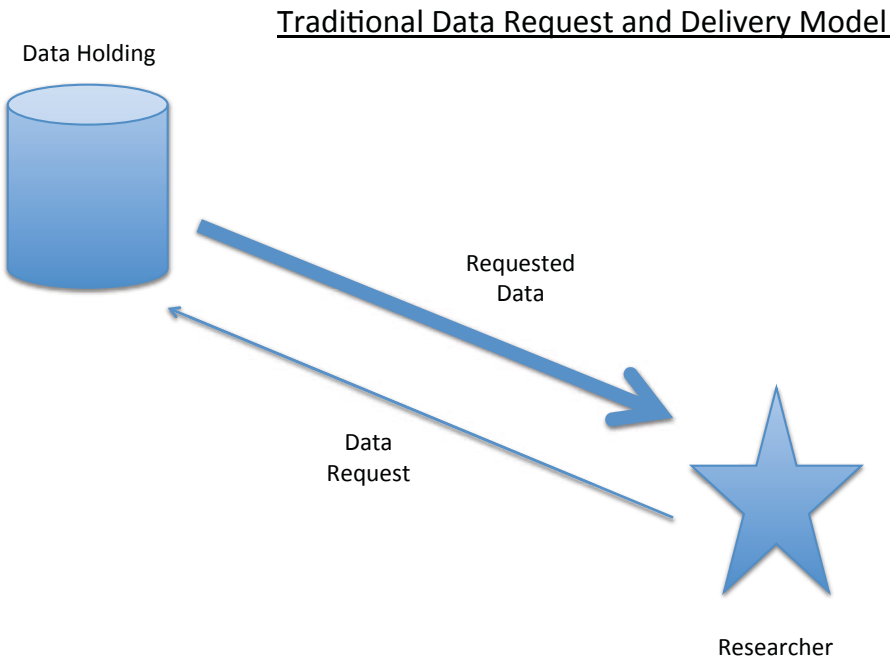

Figure 1

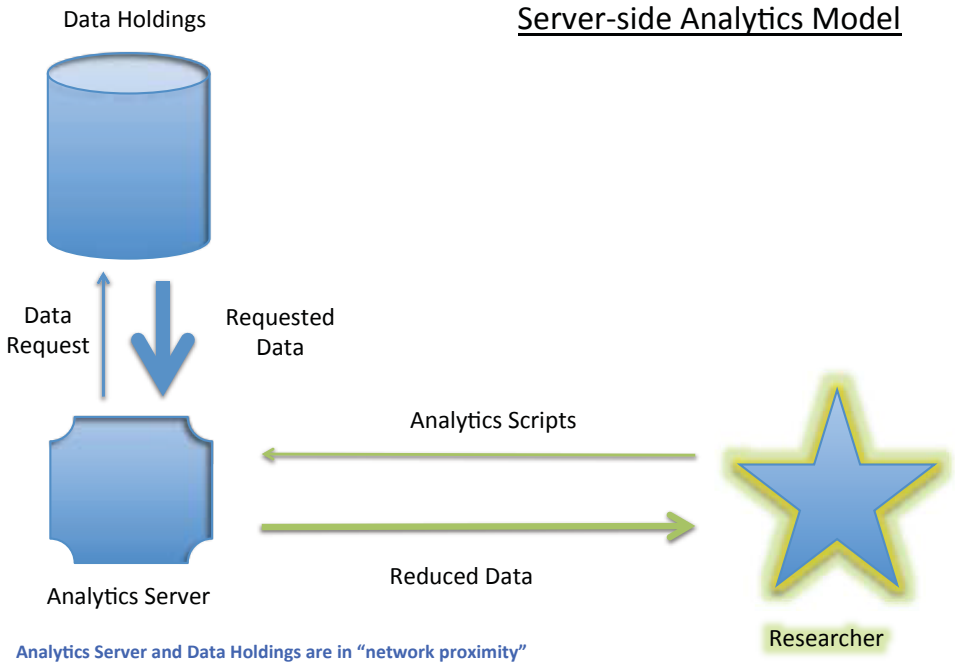

Figure 2

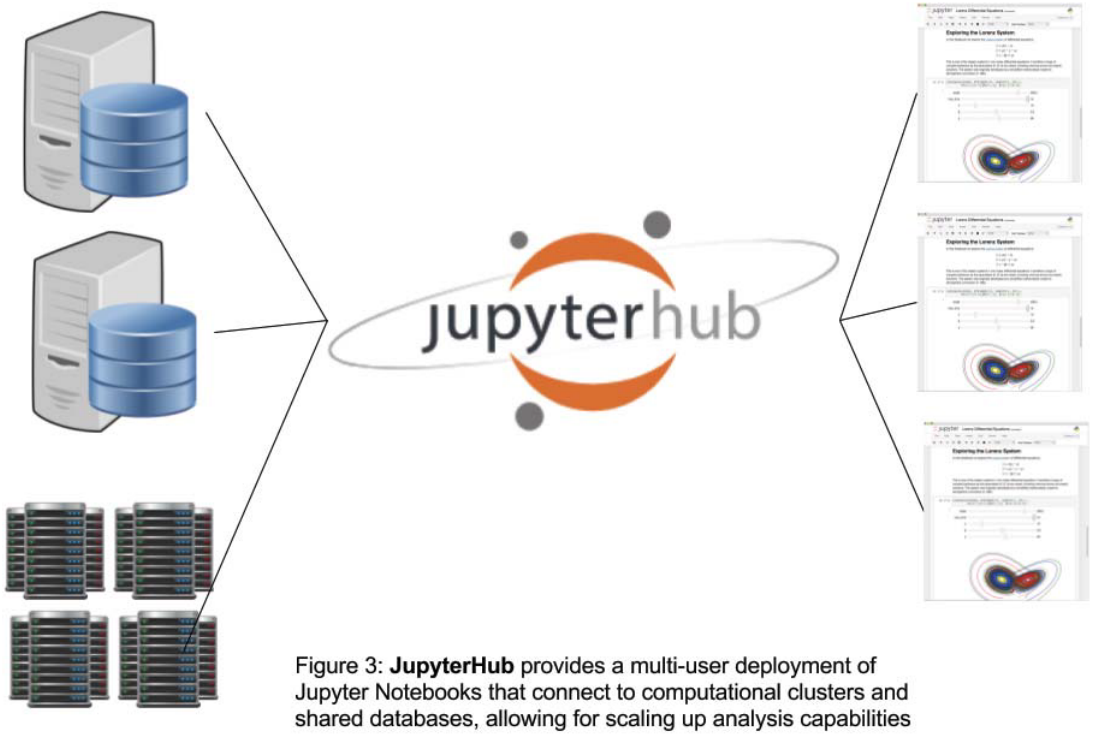

Figure 3: **JupyterHub** provides a multi-user deployment of Jupyter Notebooks that connect to computational clusters and shared databases, allowing for scaling up analysis capabilities

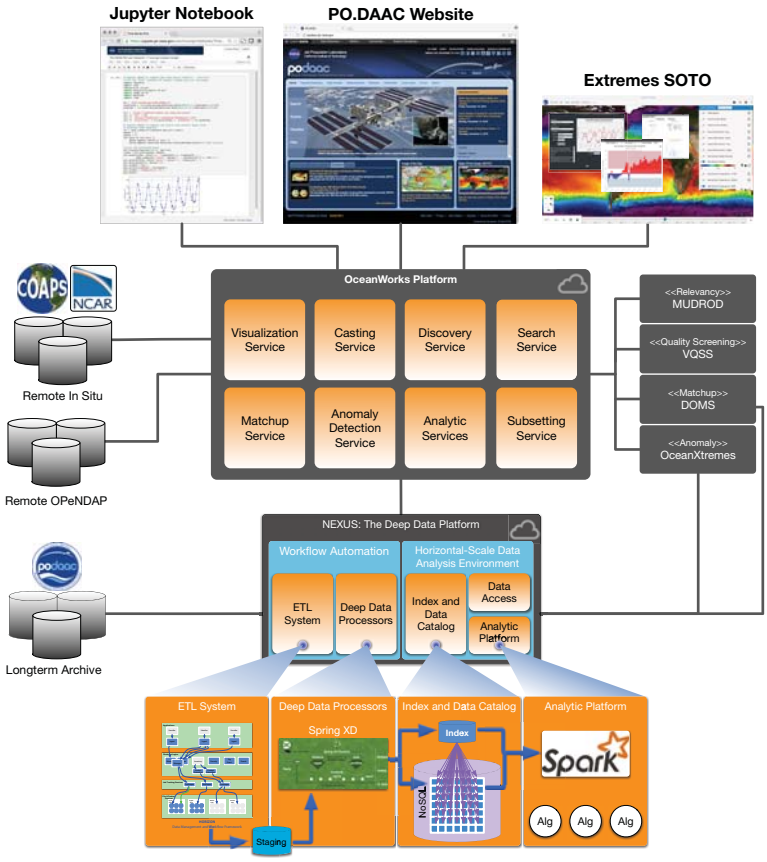

Figure 4: OceanWorks Architecture

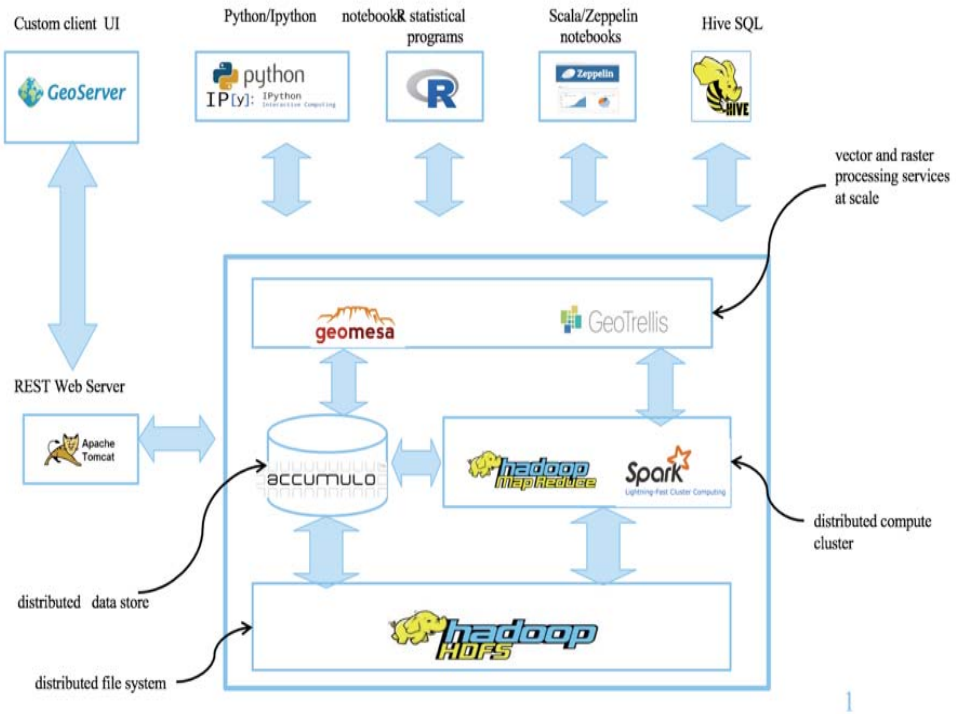

Figure 5: WaterTrek Data Analytics Architecture

	SWOT SDS	NISAR SDS	SMAP SDS	OCO-2 SDS
Daily Data Acquisition Volume	1 TB	3.25 TB	0.14 TB	0.01 TB
Daily Production Volume	15.5 TB	86 TB	0.44 TB	0.06 TB
L1 Product Latency	29 days	30 days	12 hrs.	2 days
PGE count	> 20	>10	28	17
Ancillary data types	~11	~14	@ Ops:133	@ Ops: 14
Complexity of Ops Workflow	High	Medium	High	Low

Science users will likely need to adopt new strategies for interacting with SWOT and NISAR's high-volume data

Figure 6: SWOT and NISAR Data Volumes

APPENDIX – Examples of Server-Side Analytics-type Projects

The NASA and its allied research organizations have explored with some prototyped analytics processing and, in some cases, gone on and implemented specific SSA applications. This APPENDIX presents discussions on the following projects: GrADS, MERRA/AS, OceanWorks, Western States Water Mission, SWOT and NISAR science data systems, researcher's tools for accessing and analyzing SDO data, SciServer, the NOAO DataLab and GAVIP.

1. One of the earliest examples is a package of Earth science data analysis and visualization software called the Grid Analysis and Display System (GrADS), in active development since 1989, it was modified in the early 2000s to support remote subsetting and analysis via a client-server implementation[4]. The result was called GDS - the GrADS Data Server. GDS was implemented using OPeNDAP as the communication protocol, employing http for transporting packets of information on the internet. (See: http://cola.gmu.edu/grads/grads.php) Variations GrADS and GDS are now widely deployed by NOAA and frequently accessed.

2. Also in Earth Science, the GSFC has developed a capability: MERRA/AS which is a climate analytics service specifically tailored to making MERRA[5] data more easily accessible to an expanding community of consumers, including local governments, federal agencies, and private-sector customers. This capability represents an architectural approach to climate data services that can be generalized to applications and customers beyond the traditional climate research community. Development of MERRA/AS is mature and ready for operational deployment.

3. The NASA AIST OceanWorks (https://oceanworks.jpl.nasa.gov) establishes an Integrated Data Analytic platform on the Cloud environment to enable Big Ocean Science. It focuses on technology integration, cloud infusion, advancement and maturity by bringing together several previous NASA-funded AIST and ACCESS projects as an effort to deliver a production-ready data science platform for the ocean science community. The emphasis is integration and platform building by hiding all the complexities of data management, horizontal-scaling, domain-specific technology implementations, and cloud computing architecture. User applications and services will integrate with OceanWorks through RESTful APIs and well-defined information model. OceanWorks is not your typical killer-app. It is a platform for users and external organizations to plug into a cloud-based analytics platform for on-the-fly analysis of many oceanographic measurements without having to download massive amount of data to their local computer.

Recognizing the technology solution for OceanWorks can be extended to support multidisciplinary science, the OceanWorks project has donated all of its solutions to the Apache Software Foundation into what is called the Science Data Analytics Platform (SDAP) (http://sdap.incubator.apache.org). Its goal is to create a community-supported, integrated

[4] Some of the early work to develop GrADS was funded by NASA SMD and its predecessor organization.

[5] MERRA: The Modern-Era Retrospective analysis for Research and Applications is an amalgamated data set assembled for performing global climate research. See: http://gmao.gsfc.nasa.gov/merra.

platform for big geospatial data analysis using Cloud computing technology. The SDAP currently includes many other SSA applications.

For example, the Distributed Oceanographic Matchup Service (DOMS) delivers a cloud-based matchup solution by integrating distributed in situ data hosted at JPL, NCAR, and COAPS. The project has standardized access to point-based in situ data using open source implementation of OpenSearch called the Extensible Data Gateway Environment (EDGE). DOMS translates the temporal spatial query into in situ subset requests to the external data centers. Upon receiving the sub-setting in situ data, DOMS executes its map-reduced, matchup algorithm on the cloud. The matchup result is packaged, transmitted and visualized.

Also, Mining and Utilizing Dataset Relevancy from Oceanographic Dataset (MUDROD) is a search analytic technology by continuously mining search logs from data portals. Through machine learning technology, MUDROD exposes hidden relationships between ocean datasets and dynamically adjust data ranking to show the most relevant datasets first. With data centers like the NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC) or projects like EOSDIS, which manages large collection of data, the ability to finding related data, services, and publications is a game changer.

4. Another JPL project is the Western States Water Mission (WSWM) project at is addressing the question "how much water do we have?" by understanding hydrological flows and stocks with data from space-based, airborne, and in situ measurements, and NASA's assimilation capabilities. Coupled and validated models will be utilized in a robust state estimation effort to improve the accuracy of water availability information. WSWM data and analytics provide estimates, with quantitative uncertainties, of freshwater resources, namely snow water equivalent, surface water, soil moisture, and ground water. WSWM will also estimate fluxes like evapotranspiration rates, runoff, and stream flow. Within the context of WSWM, it will be possible to demonstrate the architecture, capabilities and technologies of a scalable data analytics center enabling an interactive interface to massive hydrological data.

The data science goals are to: (1) Develop analytics that exploit a massive back-end data store to produce estimates relevant to the water management in the Western U.S.; (2) Integrate uncertainty management to enable scientific validation of model data against in situ data; (3) Leverage a high-capacity back end data store to enable timely execution of key machine learning diagnostics, such as novelty detection, on these hydrological data; (4) provide APIs that enable users to run their own computations. The data analytics platform confronts many big data and data science challenges with scalability of distributed data, data integration of heterogeneous data, uncertainty quantification from observations to derived inferences, on-demand computation for real-time data analytics, different database containers to better support interactive deep dive analysis.

5. The SWOT (https://swot.jpl.nasa.gov) and NISAR (https://nisar.jpl.nasa.gov) missions are under development and are scheduled to be launched in about 2020 and 2022, respectively. Both missions will deploy synthetic-aperture radars and will have daily data accumulations of 15 to 86 TB (see Figure 6.) which is unprecedented for NASA science missions. The science data centers (SDS) for both of these missions will probably employ cloud architectures for both processing and data storage. This could involve co-locating the traditional functions of the DAAC adjacent to the

mission's SDSes within a cloud framework. This opens opportunities to provide server-side analytics functions at the data centers and truly "Move the Analysis, not the Data."

6. There have been several efforts in the field of Heliophysics since 2000, driven by the impending flood of data coming from the Solar Dynamics Observatory (SDO). SDO generates 2TB of data each day and its data management strategy was critical to its success. SMD grants lead to the development of a set of web services that enable outside users to operate upon data hosted on a data servers (see http://www.lmsal.com). These services included data cutouts, movie generation and prototypes for remote access to more extensive processing tools.

As SDO launched, similar services were developed as part of the Helioviewer project (http://helioviewer.org) which included integration with social media services.

Two of the prime science teams (AIA and HMI) took on the task of devising schemes to deliver useful science products at reasonable data volumes. The Joint Science Operations Center (JSOC, http://jsoc.stanford.edu), where AIA and HMI data are processed, currently supports a variety of on-demand, pre-processing scripts to help reduce the volume data transfers. These include scripts to extract sub-images from the full 4kx4k images and subsample in time; to scale images and transform them into a different of coordinate systems and map projections; and to create jpeg images and compressed H.264 movies. There is also support for users to bring their analysis codes into the JSOC for execution on internal servers in the same manner as other centers of High Performance Computing (e.g. NASA/Ames' Pleaides system).

The Heliophysics Events Knowledgebase (http://www.lmsal.com/hek) was developed to try to anticipate user requests using data mining algorithms to identify interesting subsets of AIA and HMI image data as they streamed in. The HEK created a framework and schema that enabled community members to contribute algorithms that can be integrated into the science pipeline. Metadata and summary movies of the results are posted online with links to web services for extracting relevant subsets of the AIA data.

In Astronomy and Astrophysics there are several early examples of server-side analytics services.

8. SciServer, (http://www.sciserver.org) "a collaborative research environment for large-scale data-driven science," is funded by the NSF and run by The Johns Hopkins University. This project grew out of the data service activities for the Sloan Digital Sky Survey (SDSS)[6]. "SciServer is a revolutionary new approach to doing science by **bringing the analysis to the data**. SciServer consists of integrated *Tools* that work together to create a full-featured system." SciServer is proving to provide advantages not only to astronomers using SDSS data but to a wide-ranging array of activities including genomics, oceanography, soil ecology, etc. More recently the SciServer team has been making a lot of progress in extending SciServer services to medicine and materials science.

[6] The SciServer framework grew out of **SkyServer**, a website first created in 2001 for the Sloan Digital Sky Survey. The still-ongoing SDSS (http://www.sdss.org) uses a telescope in New Mexico to take images of nearly half a billion stars and galaxies, resulting in a high-resolution map of the Universe.

The MAST (web reference) archive at the Space Telescope Science Institute has been exploring some capabilities to bring the SciServer features for accessing data from JWST and its newly acquired Pan-STARS data base.

9. The National Optical Astronomy Observatory is developing Data Lab (http://datalab.noao.edu/index.php) which according their web site: "The Data Lab provides services to enable as much work as **possible close to the data**, while allowing transfer of data and results to local hardware anytime during the process. These pages contain information on the services that the Data Lab provides and the clients available to access them." "The Data Lab aims to:

- Connect users to high-value catalogs from NOAO and external sources (e.g. SDSS, Gaia) and NOAO-based images linked to catalog objects
- Enable users to discover the data that they need for their science
- Allow users to develop intuition through interaction with selected catalog and image sets
- Allow users to automate their analysis to aid discovery in large datasets"

10. GAVIP, the Gaia Added Value Interface Platform[7], is a demonstration project under development in Europe that will provide "an innovative platform within which scientists can submit and deploy code, packaged as 'Added Value Interfaces', which will be **executed close to the data**." The project recognizes that "there is an increasing need to facilitate the further use of Gaia products without needing to download large datasets." Managers from ESA's Science Operations Centre plan to incorporate GAVIP into their wider Science Exploitation and Preservation Platform.

[7] For abstract see: http://adsabs.harvard.edu/abs/2016SPIE.9913E..1VV. GAVIP documentation here: http://docs.gavip.science/ - GAVIP GitHub repositories here: https://github.com/parameterspace-ie

BDTF Recommendation for Implementing Server-side Analytics Architectures

NASA's archives of science data are growing larger and incorporating more features. In many cases efforts needed by users of these data are being hampered by the inefficiencies associated with moving large data sets from the archives across the internet into the storage local to the user. A new paradigm is beginning to appear which will permit operations on the data prior to transmission. These operations reduce the amount of data that will be transmitted. There are several terms being used to describe the new paradigm: data analytics, climate analytics, server-side analytics, etc., which all have common features – namely bring data analysis code to the data.

The BDTF has studied the case for adopting SSA implementations where needed and provides the following recommendations:

- Publish an RFI via NSPIRES to ascertain where there are centers of SSA development that could be useful for NASA-sponsored research.
- Convene workshops for the NASA science community to show off the SSA services that are under development and get feedback that could point to where new developments are needed,
- Solicit via ROSES for demonstrations or prototypes of SSAs that have the most promise of solving "log jam" situations in priority areas.
- For those projects that demonstrate potential at significantly improving the delivery of tailored processed data to the researcher, SMD should take steps to make those projects operational.

In all events members of organizations who provide SMD data archive services need to participate. It is most important that representatives, ideas, and approaches from the broader data user communities need to participate and in certain cases lead in the implementations

Sound science leadership from SMD will be needed to ensure that new SSA architectures are implemented where most needed and other proposals for lower priority implementations are reserved. The use of community peer groups and advisory working groups will be valuable in providing priority rankings for new concepts and proposals.

<u>Rationale for the recommendation:</u>
The rationale for moving to server-side analytics architectures include:

- Reading data over the internet is slower than reading data from a local file.
- Some analysis tasks require a lot of data: *"Compute the mean annual cycle from a 100-year climate model run."*
- Analysis expression is evaluated at the server, where the data reside on disk.
- Only the analytics result is delivered to the client.
- SSAs can save a lot of time if: (size of result) << (size of data operated on).

<u>Consequences for not adopting the recommendation:</u>
We are already seeing examples where data analysis communities are restricted by the bandwidth of the transmission from the data archives. This problem will get worse due to ever increasing data volumes. If we do not take action now, we will not realize the full potential for producing new science from these data sets and possibly not meeting L1 requirements.