Enabling Analytics in the Cloud for Earth Science Data

21-23 February 2018 | Annapolis, MD

Workshop Report

Workshop Committee

- Rahul Ramachandran, NASA/MSFC
- Christopher Lynnes, NASA/GSFC
- Andrew Bingham, NASA/JPL
- Brandi Quam, NASA/LaRC

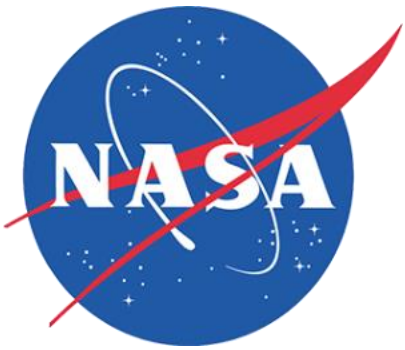

Table of Contents

Executive Summary

NASA hosted the "Enabling Analytics in the Cloud for Earth Science Data" workshop on February 21-23, 2018 in Annapolis, Maryland. The purpose of this workshop was to hold interactive discussions where providers, users, and other stakeholders could explore the convergence of three main elements in the rapidly developing world of technology: Big Data, Cloud Computing, and Analytics.

Findings and recommendations from the workshop are centered around six main themes:

1. Strategic Alignment of Cloud Analytic Efforts

The three HQ Program Offices are tackling cloud-enabled data analytics within their purview and are providing, or planning on providing, cloud environments. While they all have clearly aligned objectives and common ideas, there appears to be limited coordination between these three offices. Closer coordination between programs and awareness of others' initiatives, priorities, and investments is necessary in order to prevent duplication of effort. Maintaining a searchable portfolio of projects and services will help in discovery services/tools that *cross program boundaries.* These programs should also partner with the private sector to leverage existing and future commercial cloud analytic platforms.

2. Common Reference Architecture for Cloud Analytics

The workshop revealed that the community has yet to develop a consensus around a single architecture, as well as agree on common terms and definitions. However, one area the represented community agreed on was the need to create intermediate cloud data stores that provide views into the data archive and are optimized for fast and scalable data analytics. The adoption of a service-based architecture that

consists of data services and analytics services will alleviate the need for a common architecture. Data services should adopt and advance industry specifications that support preprocessing and access to cloud-optimized data stores. Analytics services can build upon the data services to provide methods that reduce and transform the data for the purpose of scientific exploration.

3. Analysis-Optimized Data

Workshop participants identified "Analytics Optimized Data Stores" (AODS) as a solution to Big Data Volume and Variety challenges. Participants defined AODS as data stored to minimize the need for data-wrangling and preprocessing for a large community of users that supports fast & parallel access. AODS are optimized and cost-effective storage structures for queries relevant to their users to enable iterative science analysis.

Building Analytics Optimized Data Stores to serve as building blocks for analytic tools and services and exposing APIs (web-services) is viewed as a best path forward. AODS' construction must be transparent with well-documented provenance to foster trust by the user community.

4. Discovery and Reuse of Cloud Analytics-Related Services

The cloud computing model of reuse via services needs to be extended to encompass NASA cloud analytics-related services. Every significant capability and function needs to be constructed and exposed as a service. Reuse of these capabilities will then become a simple matter of consuming the service. Establishment and use of Community Best Practices in architecting new systems/tools based on services as well as incentivizing reuse of existing services was recommended by the participants.

5. Deep Learning / Machine Learning Adoption for NASA Earth Science Data

The challenge for wider adoption of ML/DL is no longer the lack of usable algorithms, tools, or compute resources but rather the dearth of sufficient training (labeled) data in Earth science. Access to training data is required to entice DL practitioners to tackle Earth science problems. There is a need to support R&D efforts to create new strategies for generating labeled training data in Earth science, and these training data need to systematically be distributed and archived along with benchmark datasets to test algorithm performance. Use of field campaign and in-situ ground truth data as training sets also needs to be explored.

6. Cloud Analytics Adoption and Enablement Strategy

Cloud adoption by the user community is not a foregone conclusion, and efforts need to be directed toward fostering that adoption. Training mechanisms and documentation for end users to learn and understand the realm of possibility in cloud analytics should be developed. Science meetings and workshops need to be effectively utilized in order to increase awareness of cloud analytics capabilities. User-adoption lessons learned from existing cloud efforts such as ADAPT / ABoVE, AMCE, and OpenNEX need to be mined to shape the adoption strategy.

ACKNOWLEDGMENTS

Workshop Sponsorship

- Kevin Murphy, ESDS Program, NASA/HQ
- Andrew Mitchell, ESDIS Project, NASA/GSFC

Logistics and Planning

- Taylor Wright, NASA/MSFC (MIPSS)
- Jay Gentry, NASA/LaRC (BAH)
- Micheline Meyers, BAH
- Hilda Hernandez, BAH

Section 1.0: Introduction

Rapid developments in technology are changing the way data-driven research can be conducted in the Earth sciences. Improvements in sensor technologies, a wide variety of data sources, accelerated data gathering rates, and variability in accuracy and completeness of data have all led to the actualization of the **Big Data** phenomenon.

The emergence of **Cloud Computing** as a scalable compute infrastructure now available to all researchers is also changing the way scientists analyze data. These cloud infrastructures provide server-less computing, ease of use, and the ability to scale up to extremely large datasets, as well as to easily share and collaborate with others. Infrastructures are also no longer limited to hardware and provide a full suite of managed services to make Earth science data analysis-ready.

Analytics is at the core of extracting value from data in order to create new scientific knowledge and insight and thus ties Big Data to Cloud Computing. Analytics encompasses new approaches of data exploration, sophisticated statistical analysis, and Machine Learning. Researchers require new analytic tools to explore extremely large datasets interactively. These tools often rely on underlying compute infrastructure and require new architectures for data storage and organization, restructured data from traditional files to highly scalable databases, and distributed file systems. Thus, new data-parallel processing systems for analysis may require refactoring existing science algorithms. Recent breakthroughs in Machine Learning algorithms such as Deep Learning require both Big Data and the availability of scalable computing to be effective.

The "Enabling Analytics in the Cloud for Earth Science Data" workshop explored the convergence of the following three elements: Big Data, Cloud Computing, and Analytics. Several projects, supported by many organizations both internal and external to NASA, are currently working in each or all of these elements. These projects have generally been designed to meet specific goals and, unfortunately, may sometimes be duplicative in nature. Because these various projects are presented and promulgated at different conferences and within different communities, a challenge exists in better understanding the current state of the art within these three elements. This workshop enabled participants from these projects to present key information about their projects, learn about related work, and identify any existing gaps that should be addressed in the future.

Participants also had the opportunity to discuss and develop a common notional architecture (i.e., how to enable analytics in the cloud for Earth science data), where the notional architecture is an idealized description of the components, services, and interfaces required and with no assumptions about physical implementation. The workshop did not seek to converge on a specific implementation, approach, or technology; rather, it encouraged participants to think broadly in order to identify commonalities and existing gaps.

Section 2.0: Workshop Overview

The workshop brought together a diverse group of stakeholders serving unique missions to explore opportunities associated with enabling Earth science data consumers to perform data analytics in the cloud. Invited stakeholders included:

- **Engineers and Practitioners** developing data analytics solutions in the cloud in order to enable the distribution of information and gain visibility for their tools, approaches, and decision-making strategies. Practitioners ranged the entire spectrum from analytic algorithm developers to tool developers, and from analytic system architects to data system architects.
- **Researchers** cultivating current analytics applications or working with various ideas of new cloud applications that can serve as use cases for future solutions.
- **Project and Program Managers** currently leading or planning to support cloud analytics efforts.

2.1 Workshop Presentations

Workshop presentations were purposefully selected and categorized into the following topics by the Workshop Committee. These presentations are available for reference via the links provided.

2.1.1 Programmatic Overview

- Earth Science Data Systems Program Update (Kevin Murphy)
- Machine Learning in NASA's Earth Science Division (Michael Little)
- High-end Computing Program's Cloud Strategy and Plan (Tsengdar Lee)

2.1.2 Science Motivation, Drivers, and Use Cases

- Connecting Space to Village (Raymond French and Francisco Delgado)
- Strategy for Global Land Change Monitoring (Alexandra Tyukavina)

- ACCESS to Terra Fusion Cloud (Guangyu Zhao)
- Open Data Cube (Syed Rizvi)
- From HPC to the Cloud: Challenges of Large-Scale Earth Science Analytics (Daniel Duffy)

2.1.3 Analytics Algorithms and Tools

- Machine Learning based Knowledge Discovery…at Scale (Kamalika Das)
- Leveraging Geospatial Technologies for Earth Science Analytics in the Cloud (Brandi Quam and Brian Tisdale)
- Towards Earth Science Deep Learning: Applications, Platforms, and Challenges (Manil Maskey)
- Value of Deep Learning in Earth Science (Sangram Ganguly)
- STARE at the Center of Earth Science Data Interoperability (Kwo-Sen Kuo)
- ClimateSpark: An In-memory Distributed Computing Framework for Big Geospatial Data Analytics (Chaowei Phil Yang)
- Data Container Comparison Study (Chris Lynnes)
- Full Life Cycle of Data Analysis with Climate Model Diagram (Seungwon Lee)

2.1.4 Analytics Systems and Architecture

- Advanced Rapid Imaging and Analysis (ARIA)—Architecture and Lessons Learned (Hook Hua)
- Earthdata Cloud Analytics Project (Chris Lynnes)
- Plug-and-Play Analysis and Analytics of Soil Moisture Observations in the Serverless Cloud (Matt Ueckermann)
- Climate Analytics as a Service (John Schnase)
- Pangeo - A community-driven effort for Big Data geoscience on HPC and Cloud (Ryan Abernathey)
- Open Source Architecture for Fast Earth Science Data Analysis on the Cloud (Thomas Huang)
- Giovanni in the Cloud: Earth Science Data Exploration in Amazon Web Services (Mahabaleshwara Hegde)

- Can a Machine Win a Nobel Prize? Implications for Cloud Analytics (Victor Pankratius)

2.1.5 Data Systems Architecture

- Cumulus - Data Ingest/Archive (Katie Baynes)
- Cloud Processing Architectures (Dan Pilone)
- Enabling End User Analytics at Scale (Chris Stoner)

2.2 Workshop Notes

Twenty-six pages of notes captured valuable discussions during the presentations and through the breakout sessions.

Insights gleaned from the workshop helped shape actionable recommendations for program sponsors and key stakeholders. These recommendations can form the basis of strategic next steps to address gaps in cloud analytics capabilities. Findings and recommendations from the workshop are centered around six main themes:

1. Strategic Alignment of Cloud Analytic Efforts
2. Common Reference Architecture for Cloud Analytics
3. Analysis-Optimized Data
4. Discovery and Reuse of Cloud Analytics-Related Services
5. Deep Learning / Machine Learning Adoption for NASA Earth Science Data
6. Cloud Analytics Adoption and Enablement Strategy

Section 3.0: Findings and Recommendations

3.1 Strategic Alignment of Cloud Analytic Efforts

Findings

The first session of the workshop focused on Programmatic Overviews, with presentations from three key NASA HQ Programs: Kevin Murphy (Earth Science Data Systems (ESDS)), Michael Little (Advanced Information System Technology (AIST)), and Tsengdar Lee (High End Computing (HEC)). These presentations set the stage for workshop participants and provided programmatic context and background for each program.

Kevin Murphy outlined ESDS's Earthdata Cloud 2021 strategy: (1) to migrate NASA's Earth science data from individual storage systems at each of the twelve NASA Distributed Active Archive Centers (DAACs) to the cloud, and (2) to run DAAC operations using a NASA-managed cloud environment in the near future.

Michael Little presented a concept of Analytic Centers that could be instantiated or spun down with little effort, low cost, and little required overhead when not in use. Under this concept, the specific analytic center can integrate existing algorithms/tools and rapidly evolve as components change over time. The AIST Program provides AMCE, a managed cloud environment, to its stakeholders. AMCE is complementary to the Earthdata Cloud 2021 strategy and focuses on enabling innovation. AMCE, which operates outside NASA's firewall, enables users to rapidly utilize cloud capabilities without many of the operational security restrictions inherent in NASA's firewall. Little also highlighted lessons learned from a recent Data Mining Workshop: (1) data grooming takes a long time, and (2) NASA Earth science needs to find a way to trust AI and ML.

Tsgendar Lee presented HEC's cloud strategy with a focus on model simulations and model data. Gearing up for the Big Data Paradigm, his program focuses on two

categories of data-driven/data-intensive science: (1) data assimilation and inverse problems, and (2) traditional statistics, informatics, and analytics approaches. His program is moving towards a hybrid cloud architecture with the ability to burst into commercial clouds depending on science needs. His program has two different cloud implementations, ADAPT (Advanced Data Analytics Platform) and NEX (NASA Earth Exchange), that couple computing, data, and analysis tools.

The three HQ Program Offices are tackling cloud-enabled data analytics within their purview and are planning to provide, or already provide, cloud environments; however, there appears to be limited coordination between these offices although there are clearly aligned objectives, common ideas, and intent.

In addition, several science and application use cases presented at the workshop leveraged commercial cloud analytic platforms, such as Google Earth Engine (GEE) and ArcGIS (Esri SaaS), for their analytic needs. These platforms offer many benefits and are tools of preference for many consumers of cloud analytics; these should be considered in future strategic planning efforts.

Recommendations

1. The three HQ Program Offices (i.e., ESDS Program, AIST Program, and HEC Program) should prevent duplication of effort via closer coordination between programs and awareness by each program of the other programs' initiatives, priorities, and investments, which would result in common threads of development. It would be useful for each program to communicate more frequently on program goals and funded activities and to develop a process for technical exchange for projects of mutual benefit. Suggestions include holding a focused Annual Workshop as a mechanism for technical information exchange and leveraging existing events such as ESIP, ESTF, or ESDSWG.
2. The ESDS, AIST, and HEC Programs should maintain a searchable portfolio of projects and services for anyone (ESDIS, development teams, Program Managers, Program Executives) to discover services/tools that *cross program*

boundaries (rather than, or in addition to, individual program portfolios), increase infusion and collaboration potential, and prevent duplication of effort.

3. The ESDS, AIST, and HEC Programs should develop a strategy for partnering with the private sector to utilize existing and future commercial cloud analytic platforms.

3.2 Common Reference Architecture for Cloud Analytics

Findings

Chris Lynnes presented a high-level notional architecture that defines and describes the major components within a cloud-base data analytics system and their interactions. Dan Pilone and Victor Pankratius also presented notional architectures that contain similar elements but differed significantly enough to suggest that the community has yet to develop a consensus around a single architecture, as well as agree upon common terms and definitions. These gaps were illustrated in the presentations from developers of domain-specific data analytics systems (e.g., Hegde (multidiscipline), Huang (oceans), Lee (climate), Schnase (climate), and Yang (multidiscipline)), who demonstrated different architectural views and used different terms and definitions.

One area in which the notional architects all agreed, however, was the need to create data stores in the cloud that provide views into the data archive and are optimized for fast and scalable data analytics. Depending on the presenter, the term for these data stores varied (e.g., "Analysis Ready Data (ARDs)", "Analysis Optimized Data Stores (AODS)", "Archive of Convenience", "Data Chunks", etc.). Because the structure of these data stores has significant impact on cost and performance, we believe this is a topic that needs urgent investigation (see Section 3.3).

Recommendations

The workshop committee recognizes that data analytic systems will emerge to serve specific communities, likely aligned with domains of the DAACs and/or geophysical measurement. They also recognize that technologies that support cloud-based data analytics are changing rapidly, which is driving early adopters to explore different approaches. However, NASA Earth Science should balance innovation with the need to evolve via a solution that meets operational demands. In particular, it should guard against the innovation of monolithic data analytic systems that serve a small, specialized set of users who are abstracted away from the data by data manipulation software services.

1. NASA Earth Sciences (HQ, ESDIS, and DAACs) should work toward the adoption of a service-based architecture that consists of data services and analytics services. Data services should adopt and advance industry specifications that support preprocessing and access to cloud-optimized data stores. Collaboration with external data partners is paramount to support interoperability with other data holdings through the adoption of common APIs, terms, and definitions. Analytics services build on the data services to provide methods that reduce and transform the data for the purpose of scientific exploration. These services should be developed within a collaborative NASA ecosystem that supports innovation, incentivizes reuse of services, and strives towards a common set of vocabularies.
2. NASA's ESDS Program should establish a working group (or extend the charter of an existing working group) charged with developing the specifications of the data services needed for data preprocessing and data access of cloud-optimized data stores. Where possible, the specifications should be approved and adopted by our non-NASA partners to support interoperability.
3. NASA Earth Sciences (HQ, ESDIS, and DAACs) should establish an ecosystem of technologists and software developers focused on developing services for analyzing different types of data. These services should be used by third-parties to create GUIs that support domain-specific communities.

3.3 Analysis-Optimized Data

Findings

Like all Big Data problems, Earth Science datasets from satellites, in-situ sensors, and model runs suffer from the four Vs challenge: Volume, Variety, Velocity, and Veracity. Volume and Variety are particularly difficult to approach with users of these data. High data volume stresses the end-to-end system by demanding efficient network bandwidth to download data to a local analysis capability, which in turn needs sufficient storage and compute to execute analysis. Variety of data requires reprocessing by users and is a significant, sometimes overwhelming, time sink. While common formats and metadata standards have helped, users must still invest significant time to understand the structure of the data and to condition the data, such as align them temporally and spatially, apply corrections, and filter based on data quality.

Cloud computing helps with the Big Data Volume challenge by reducing costly downloads through analysis "in place", providing theoretically unlimited compute and storage. However, to take full advantage of cloud processing, data must often be reorganized for optimal performance.

On the other hand, the preprocessing challenge is driven mostly by Variety, the virtual co-location of datasets in a cloud-based archive may foster collaboration and sharing within communities around the datasets. The collaboration and sharing can range from both preprocessing methods and the preprocessed data themselves. Several presentations applied localized preprocessing techniques in order to assemble the data into common structures, supporting fast parallelized computation and inter-comparison across datasets.

This challenge led to discussion of the need for "Analytics Optimized Data Stores" (AODS) to help address both the Volume and Variety challenges at the same time. Workshop participants defined AODS as data stored in a fashion that:

1. Minimizes the need for data-wrangling and preprocessing for a large community of users
2. Uses cloud-native storage forms to support fast & parallel access
3. Utilizes optimized storage structures for queries relevant to their users in order to enable iterative science analysis
4. Exploits cost-effective, affordable (including egress / transfer costs), and sustainable methods, and can incorporate non-NASA data from other agencies, nations, and the user community at large.

Presenters also referred to these structures by different names: data cubes, analysis ready datasets, cloud optimized data, data chunks, etc. Some data structures were presented to users through common APIs (e.g., Xarray) or custom APIs. Others did not expose the APIs, opting to present only a GUI instead. Another finding was that Jupyter notebooks were used widely across the community to interact with the APIs and facilitate interactive science.

Recommendation

1. NASA should work toward Analytics-Optimized Data Stores for data to serve as building blocks for analytic tools and services, with a focus on building and exposing APIs (web-services). Construction of AODS must be transparent with well-documented provenance to foster trust by the user community. The refined AODS definition and accompanying Extract-Transfer-Load/Preprocessing services should be developed with input from a survey of analysis use cases to identify common preprocessing operations needed to support the analyses. The use cases can be categorized by data characteristics (spatial ROI, temporal coverage, swath/grid/other), transformation operations, software languages/packages used, target user communities, and target analysis types (e.g., time series analysis, uncertainty quantification, statistical

analyses). Pilot studies should be undertaken to identify and prototype a framework of techniques and software for the translation of archived data to AODS; pilot results can help optimize data format and structure for cloud storage.

3.4 Discovery and Reuse of Cloud Analytics-Related Services

Findings

The fields of data analytics and cloud computing are both evolving rapidly. It is thus essential to enable rapid infusion of technological advances and sharing of analytics algorithms. Earth scientists desperately need scalable data science tools and are willing to help develop and use them, given appropriate incentives and assistance. In general, reuse of computer capabilities is valued by those who have the ability to utilize them effectively; however, reuse is not nearly as prevalent as it should be due to the difficulty in both producing software that is easily consumed by external parties and understanding how software can be utilized for purposes other than those for which it was originally produced.

Cloud computing has changed the infusion/reuse equation by packaging capabilities as services, extending from the bottom of the stack (Metal-as-a-Service) to the top of the stack (Software-as-a-Service). That is, capabilities are leveraged by invoking web-based (typically) services. This ability to leverage capabilities fosters decoupled architectures that are largely constituted to consume services via well-exposed service APIs, making the consumer side of reuse possible. At the same time, cloud computing services also provide an architectural pattern demonstrating to software producers how to make their software capabilities easy to reuse. This cloud computing model of reuse via services can be extended to encompass NASA cloud analytics-related services. In a cloud model, every significant capability is constructed and exposed as a service. Therefore, reuse of capabilities becomes a simple matter of consuming the service.

Recommendations

Architectural recommendations cluster around a theme of exploiting standards and patterns, with brokers to bridge divides among standards and a discovery element to tie resources together. Due to the human factors involved, particularly with respect to reuse, these recommendations address technical and sociological or programmatic issues.

1. The three HQ Program Offices (i.e., ESDS Program, AIST Program, and HEC Program) should establish and utilize a Community Best Practices to architect new systems/tools and incentivize reuse.
2. Leveraging existing and evolving standards where possible, EOSDIS Standards Office should work with the Earth Science Data Systems community to establish standard extensions for common Application Program Interfaces (APIs) and information models for service invocation. This provision should be done in concert with industry, international, and inter-agency partners. Further, there is a need to design and document patterns for micro-services (i.e., atomic services that perform only one task but can be connected to other micro-services to implement complex workflows). To handle the impedance of mismatches between different standards and customer services, service brokers can be utilized. The discovery element includes optimizing existing analytics-related resources for both application developers and science end-users. These resources include Services, Libraries, and Containers, as well as Workflows that utilize and demonstrate services and Jupyter notebooks that demonstrate services or implement workflows.
    a. NASA programs should consider these resources for inclusion as success criteria in future solicitations for programs such as ACCESS, AIST, and CMAC. NASA should also encourage community working groups to further define and promulgate these practices.
3. ESDS should establish incentives for reuse, for both producers and consumers, such as inclusion of reuse requirements in NASA Requests for Proposals and development processes, patterns, and other aids in order to make it easier for scientists to contribute to the future analytics capabilities.

4. NASA HQ and ESDIS should establish a small team of technologists to survey both Science & Technology developments and emerging technologies external to NASA and/or beyond the NASA ESDS Program, such as: community developments (e.g., Pangeo), industry developments (e.g., AWS OnEarth, Tech Radar), and other technologies of use by both cloud analytics developers and end users.

3.5 Deep Learning / Machine Learning Adoption for NASA Earth Science Data

Findings

The session on Algorithms and Tools had several interesting presentations focused on Machine Learning/Deep Learning (ML/DL). The presentations covered a range of Earth science applications and utilized a wide spectrum of algorithms including symbolic regression, scalable graph-based analysis, and DL networks. The cloud enables collocated data and provides on-demand, specialized computation resources such as GPU clusters, which make ML/DL an attractive solution for large and complex problems. These presentations demonstrated that deep learning clearly has tremendous potential to enable new science and to build new applications. New services, such as SageMaker (Amazon Web Services' ML-as-a-service), bring the power of deep learning to data, lowering the barrier to entry to utilize existing ML/DL tools.

The challenge for wider adoption of ML/DL is no longer the lack of usable algorithms, tools, or compute resources, but rather the dearth of sufficient training data in Earth science. Access to training data is required to entice DL practitioners to tackle Earth science problems. However, manually creating labeled training data is still a bottleneck and new strategies to increase training size need to be explored. One potential opportunity for NASA is to better leverage its many field campaigns and in-situ ground truth data to support the development of training data. These datasets are often of high fidelity and reliability; however, they are difficult to discover, navigate, and use, particularly for DL practitioners. This problem represents

an opportunity for NASA ESDS to add new value to campaign and in-situ data by making them easier for DL practitioners to find and use.

Recommendations

1. NASA HQ should provide support to R&D efforts that evaluate new strategies to create labeled training data in Earth science. Some possible strategies are data augmentation, transfer learning, permutation invariance, data programming, citizen science, and active learning.
2. NASA HQ should develop an approach to systematically create, distribute, and archive training/labeled data. Publication of these data holdings across both the Earth science and the computer science communities would broaden the use of deep learning techniques in Earth science.
3. The three HQ Program Offices (i.e., ESDS Program, AIST Program, and HEC Program) should develop community-wide benchmark datasets to aid in assessing new methods' performance, cost, and utility.
4. NASA HQ should explore the potential of using field campaign and in-situ ground truth data to develop training sets. Investigation into the preprocessing is required to re-structure these datasets to enable the use of ML/DL methods.

3.6 Cloud Analytics Adoption and Enablement Strategy

Findings

A consensus emerged from workshop discussions that cloud adoption by the user community is not a foregone conclusion, and that some effort needs to go toward fostering that adoption. However, there are indicators, such as the ACSI Survey results, that a substantial portion of the end-user community is open to (or already are) trying to use cloud computing for their data analyses.

Different strategies surfaced, ranging from hiding the cloud infrastructure via existing tools (such as Giovanni and SERVIR) to offering training to users on how to get

started in cloud computing. The diversity of user types in terms of novelty tolerance, ties to legacy code, and application or research area suggests that a combination of strategies will likely be necessary.

Several examples of tools were demonstrated and/or discussed during the workshop: C. Stoner demonstrated a CloudFormation template that could be used in conjunction with provisioned scripts and the Sentinel toolbox to perform Radiometric and Terrain Correction of SAR data. Several participants also pointed to Jupyter notebooks as particularly suitable for demonstrating the use of cloud analytics from simple service utilization to complex science workflows.

Recommendations

1. NASA HQ, ESDIS, and the DAACs should develop training and documentation for end-users that (a) explains the realm of possibility in cloud analytics and (b) bridges knowledge and confidence gaps by developing and sharing data recipes, cloud analytics use cases, videos, recorded webinars, and in-person seminars.
2. The three HQ Program Offices (i.e., ESDS Program, AIST Program, and HEC Program) should conduct science meetings and workshops in order to increase awareness of cloud analytics capabilities for science use. Also, each office should increase the presence of NASA's cloud practitioners at relevant science-driven technology meetings in order to make connections with potential early adopter scientists.
3. The ESDS, AIST, and HEC Programs should investigate user adoption models in existing cloud efforts such as ADAPT / ABoVE, AMCE, and OpenNEX for lessons learned. Together, each office should devise metrics to measure the end-user adoption rate of cloud computing for analysis of Earth science data and to suggest directions for improvements in both cloud adoption and overall support for cloud analytics.

Appendix A: Acronym List

ABoVE	Artic-Boreal Vulnerability Experiment
ACCESS	Advancing Collaborative Connections for Earth System Science
ACSI	American Customer Satisfaction Index
ADAPT	Advanced Data Analytics Platform
AIST (Program)	Advanced Information System Technology (Program)
AMCE	AIST Managed Cloud Environment
AODS	Analytics Optimized Data Stores
API	Application Programming Interface
ARC	Ames Research Center
ARD	Analysis Ready Data
ARIA	Advanced Rapid Imaging and Analysis
ASDC	Atmospheric Science Data Center
ASF	Alaska Satellite Facility
AWS	Amazon Web Services
BAH	Booz Allen Hamilton
CMAC	Computational Modeling Algorithms and Cyberinfrastructure
CEOS	Committee on Earth Observation Satellites
DAAC	Distributed Active Archive Center
DL	Deep Learning
ESDS (Program)	Earth Science Data Systems (Program)
ESDIS	Earth Science Data and Information System
ESDSWG	Earth Science Data System Working Groups
ESIP	Earth Science Information Partners
ESTF	Earth Science Technology Forum
ESTO	Earth Science Technology Office
GEE	Google Earth Engine
GES DISC	Goddard Earth Sciences Data and Information Services Center
GMU	George Mason University

GSFC	Goddard Space Flight Center
GUI	Graphical User Interface
HEC (Program)	High End Computing (Program)
HPC	High Performance Computing
HQ	(NASA) Headquarters
JPL	Jet Propulsion Laboratory
LaRC	Langley Research Center
MIPSS	Marshall-Integrated Programmatic Support Services
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MSFC	Marshall Space Flight Center
NASA	National Aeronautics and Space Administration
NCCS	NASA Center for Climate Simulation
NEX	NASA Earth Exchange
OpenNEX	Open Nasa Earth Exchange
R&D	Research and Development
ROI	Return on Investment
SaaS	Software as a Service
SAR	Synthetic Aperture Radar
SERVIR	Mesoamerican Regional Visualization and Monitoring System (A joint venture between NASA and the U.S. Agency for International Development)
STARE	SpatioTemporal Adaptive Resolution Encoding
UMD	University of Maryland

Appendix B: Participant List

Name	Organization/Affiliation	Role
Chris Lynnes	ESDIS	Program Committee/Presenter/Facilitator
Chris Stoner	Univ. of Alaska/ASF	Presenter
Dan Pilone	Element84	Presenter
Guangyu Zhao	Univ. of Illinois	Presenter
Mahabaleshwara Hegde	GES DISC	Presenter
Hook Hua	JPL	Presenter
Katie Baynes	GSFC	Presenter
Phil Yang	GMU	Presenter
Seungwon Lee	JPL	Presenter
Thomas Huang	JPL	Presenter
Syed Rizvi	CEOS	Presenter
Sanjay Gowda	CEOS	Participant/Co-Presenter
Matt Ueckermann	Creare, Inc.	Presenter
Kwo-Sen Kuo	GSFC	Presenter
Victor Pankratius	MIT	Presenter
Kamalika Das	ARC	Presenter
Sangram Ganguly	NEX	Presenter
Manil Maskey	MSFC	Presenter
Ryan Abernathey	Columbia University	Presenter

Alexandra (Sasha) Tyukavina	Univ. of Maryland	Presenter
John Schnase	GSFC	Presenter
Brandi Quam	HQ	Program Committee/Presenter/Facilitator
Brian Tisdale	ASDC	Participant/Co-Presenter
Francisco Delgado	SERVIR	Presenter
Mike Little	ESTO	Stakeholder/Presenter
Tsengdar Lee	HQ	Stakeholder/Presenter
Kevin Murphy	HQ	Stakeholder/Presenter/Sponsor
Cara Leckey	LaRC	Stakeholder
Benhan Jai	JPL	Stakeholder
Andrew Mitchell	GSFC	Stakeholder/Sponsor
Amanda Whitehurst	HQ	Stakeholder
Mark McInerney	ESDIS	Stakeholder
Raymond French	SERVIR	Stakeholder/Co-Presenter
Daniel Duffy	NCCS	Stakeholder/Presenter
Gabriel Borroni	HQ	Stakeholder
Alfreda Hall	HQ	Stakeholder
Rahul Ramachandran	MSFC	Program Committee/Facilitator
Andy Bingham	JPL	Program Committee/Facilitator
Justin Rice	GSFC	Participant/Tech Support

Jay Gentry	ASDC	Participant/Note-taker/Report Writing
Taylor Wright	MSFC	Coordination/Logistics/Time Keeper
Micheline Meyers	BAH	Logistics/Contract/Check-in