

Big Data and Earth Science : Opportunities for Science and Industry

Dr. Edward J. Kearns
National Oceanic and Atmospheric Administration (NOAA)

Committee on Earth Science and Applications from Space
National Academies of Science 30 March 2016



Outline

- What's the problem(s) we are trying to solve? Some Big Data Challenges
- Origins and progress of the Big Data Partnership research activity at NOAA
- Some lessons learned about the Challenges from the recent Big Data activities
- Questions and Discussion



Challenges : Preservation

- Collection and preservation of data?
 - Rapidly increasing volume is not the problem...
 - Supporting the labor necessary to steward and curate collections of data is a problem.
 - Recent advances in standards, formats, descriptions are very helpful.
 - Data security in Earth Sciences to be a future problem?
- Access to the data?
- Utilization of the data?

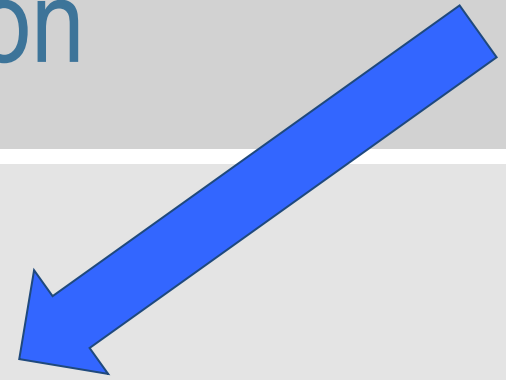


Challenges : Access

- Collection and preservation of data?
- **Discovery of and access to the data?**
 - Access for “designated communities” is probably sufficient for their research
 - Network bandwidth is a problem -- if we’re going to keep moving the data when we need to use them.
 - Discovery and access outside of a designated community is a problem. Catalogs? Free text search?
- Utilization of the data?

Challenges : Utilization

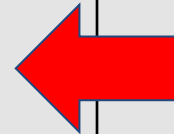
- Collection and preservation of data?
- Access to the data?
- Utilization of the data?
 - The true Big Data problem is about ease of use.
 - Applications require expert interpretation.
 - Big Data applications across fields or industries require diverse expertise – how is this expertise provided?
 - Open Data versus Proprietary Data issues: do the data or services hold the most value?
 - Pay for data? Or pay for access & services?
 - Value of curated collections, e.g. Music Services
 - Growth of platforms to use data to answer questions



National Centers for Environmental Information's

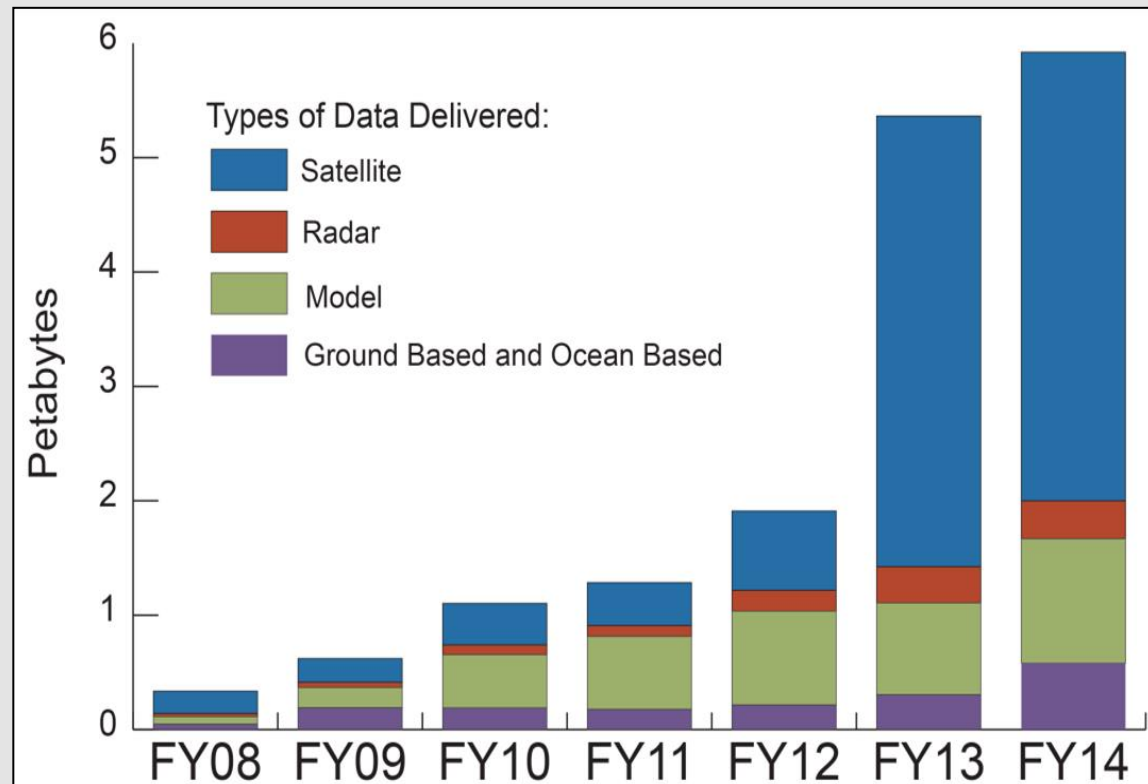
User Profile : Most want “The Answer”

Fraction (%)	Typical User	Data or Info Need	Preferred Format	Access Volume	Access Frequency
70	General business, media, public	Qualitative	Point-and-click, graphics, assessments	Low	High
15	Researchers, business consultants	Quantitative	Digital downloads	High	Low
15	Value-added Providers (database scrapers)	Quantitative	Digital downloads machine to machine	Low	High



Accelerating User Demand for NOAA data

- NOAA's National Centers for Environmental Information (NCEI) alone is now serving >8 PB of data annually
- Servicing over 20,000 personal contacts across many sectors
- 2.6 billion web hits in FY14 with 19 million users
- Significant load on NOAA infrastructure (\$\$\$)



Intro to NOAA Big Data Partnership (BDP)

1. NOAA has a lot of data, which is expensive to store and disseminate.
2. Due to a variety of accessibility issues, much of NOAA's environmental data are also **under-utilized** – especially beyond the expert community.
3. **There is untapped economic value in all that data.**
4. That value can be leveraged to improve accessibility and pay for staging of that data on the public cloud, where people and organizations of all kinds can innovate as part of a market ecosystem.



April 2015: Partnership Announcement

Unusual, no-net-cost proposition

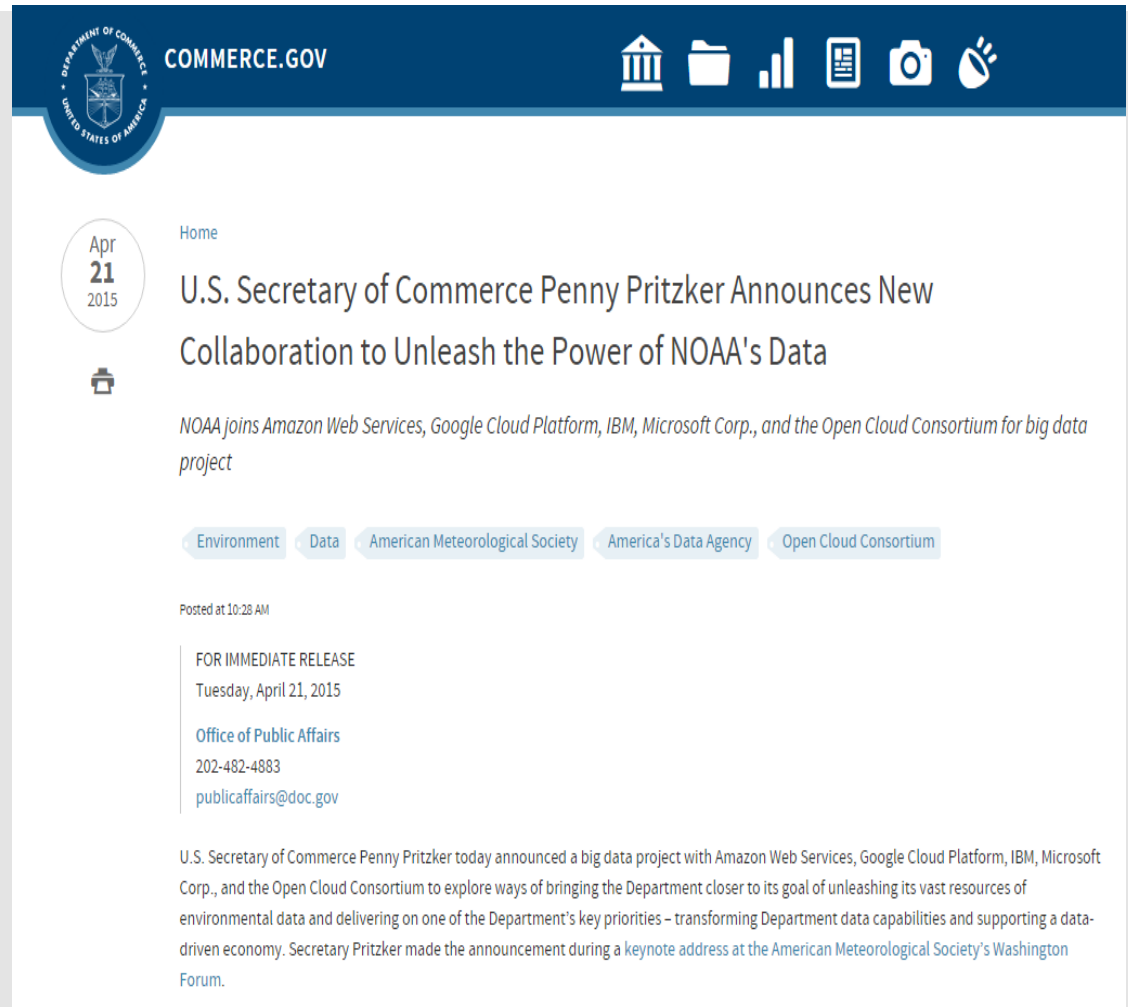
- Could the value inherent in NOAA's datasets support the cost of their distribution?

Enthusiastic, cross-industry response

- Interest from 200 companies
- 70+ responses to NOAA's "request for information"

NOAA is using Cooperative Research and Development Agreements (CRADAs), not typical acquisitions

Selection/announcement made on 21 Apr 2015



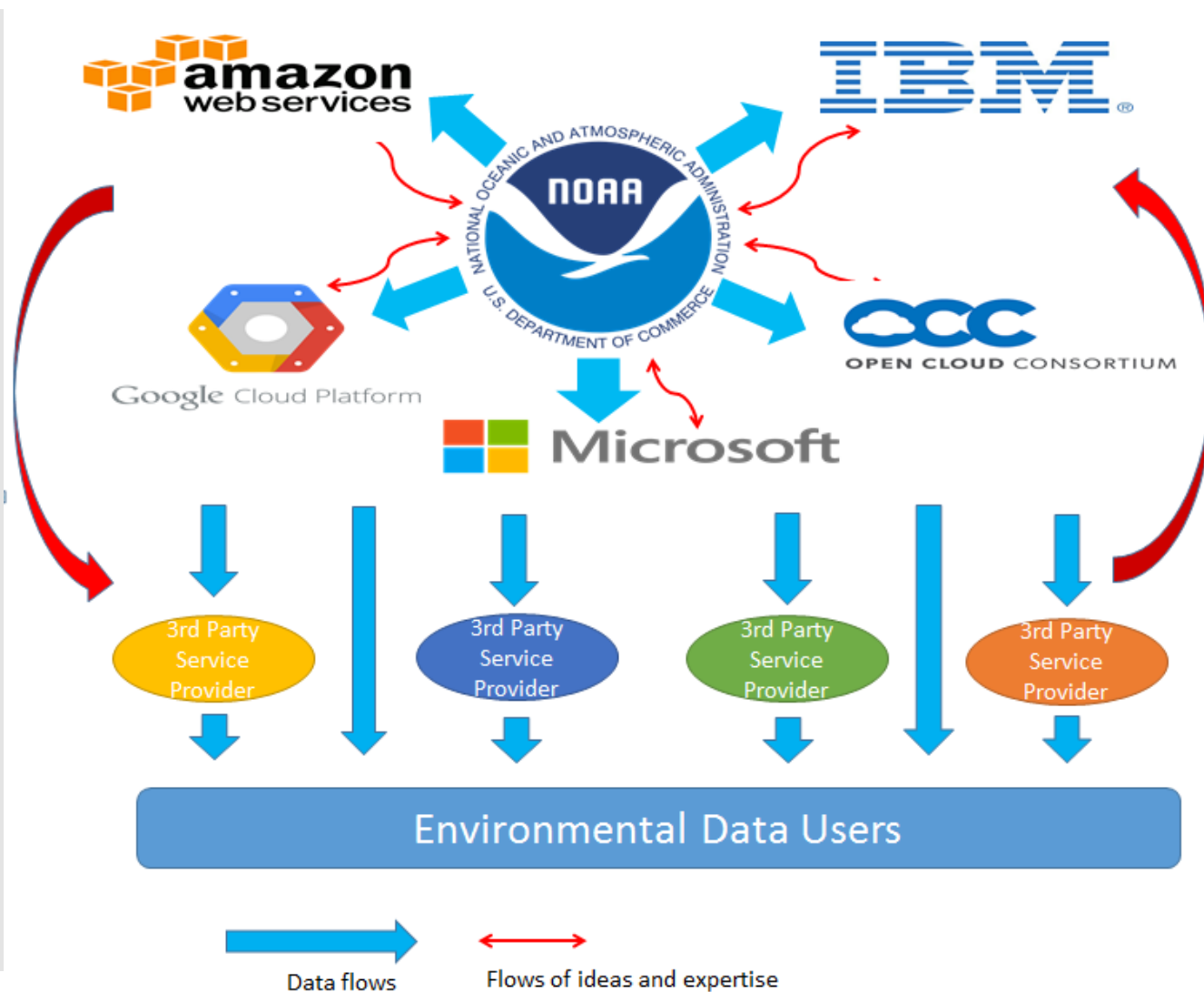
The screenshot shows the Commerce.gov website header with the Department of Commerce seal and navigation icons. The main content area features a date stamp for April 21, 2015, and a headline: "U.S. Secretary of Commerce Penny Pritzker Announces New Collaboration to Unleash the Power of NOAA's Data". Below the headline is a sub-headline: "NOAA joins Amazon Web Services, Google Cloud Platform, IBM, Microsoft Corp., and the Open Cloud Consortium for big data project". A row of tags includes "Environment", "Data", "American Meteorological Society", "America's Data Agency", and "Open Cloud Consortium". The post is dated "Posted at 10:28 AM" and includes contact information for the Office of Public Affairs: "FOR IMMEDIATE RELEASE", "Tuesday, April 21, 2015", "Office of Public Affairs", "202-482-4883", and "publicaffairs@doc.gov". The body text states: "U.S. Secretary of Commerce Penny Pritzker today announced a big data project with Amazon Web Services, Google Cloud Platform, IBM, Microsoft Corp., and the Open Cloud Consortium to explore ways of bringing the Department closer to its goal of unleashing its vast resources of environmental data and delivering on one of the Department's key priorities – transforming Department data capabilities and supporting a data-driven economy. Secretary Pritzker made the announcement during a keynote address at the American Meteorological Society's Washington Forum."

BDP CRADA Collaborators

- IaaS companies to serve as project anchors
- These CRADA “Collaborators” are nuclei for data alliances and markets
- Members of industry, research, and academia may join these alliances
- NOAA receives data requests from the Collaborators...

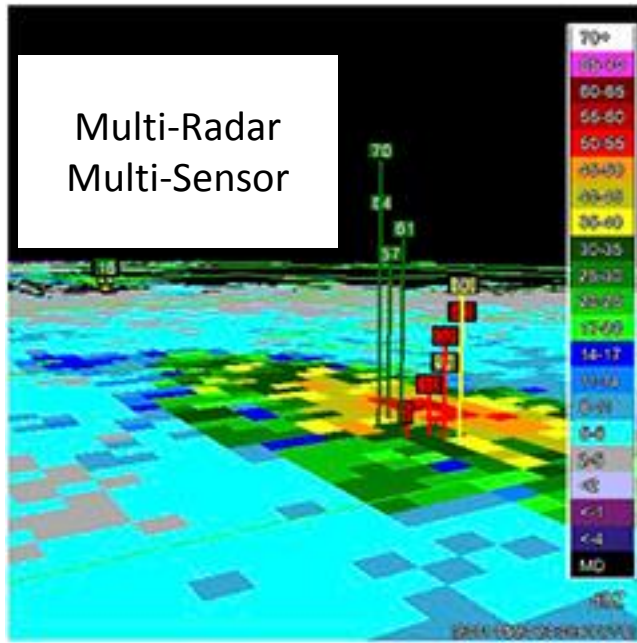


Conceptual Relationship for BDP

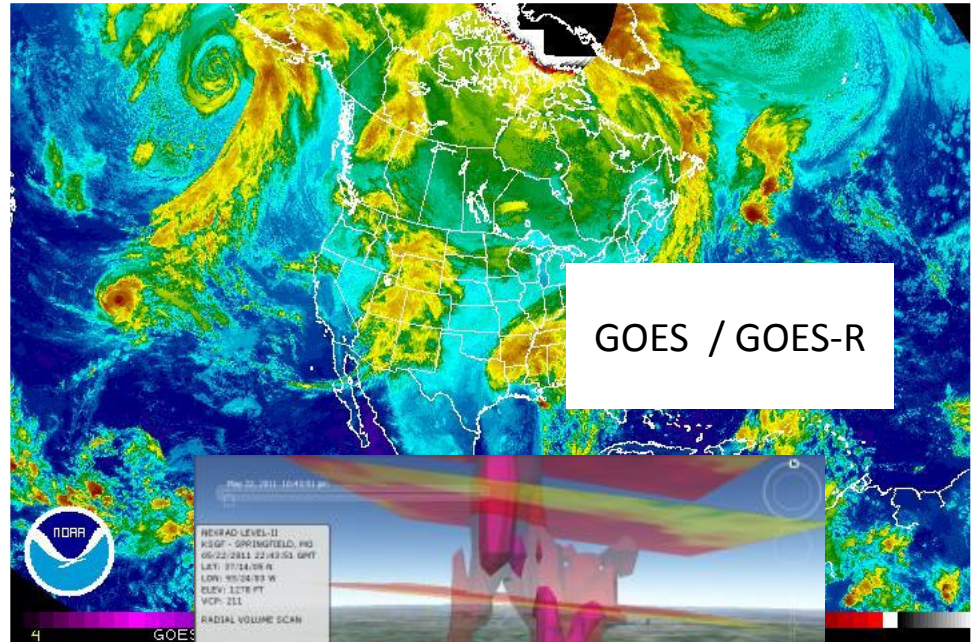


Requested Initial Data Included...

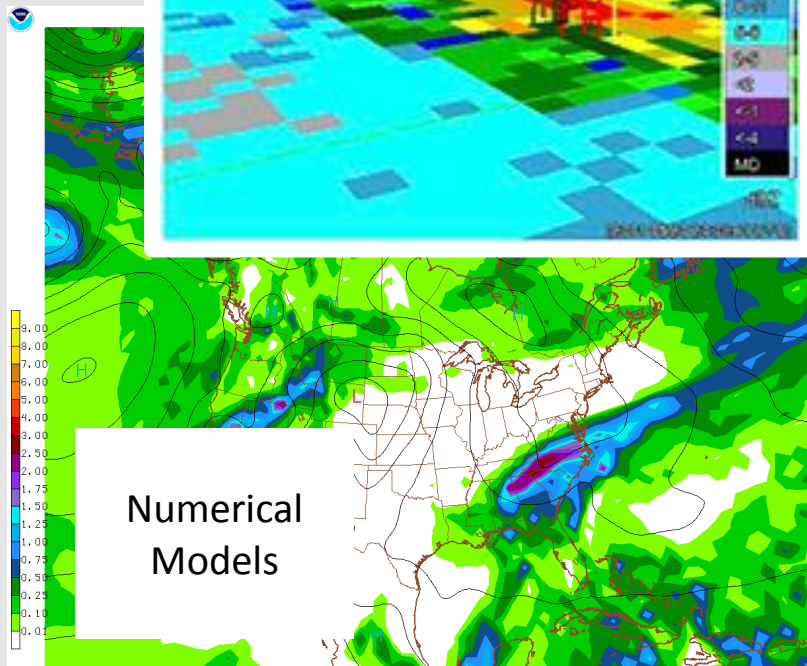
Multi-Radar
Multi-Sensor



GOES / GOES-R



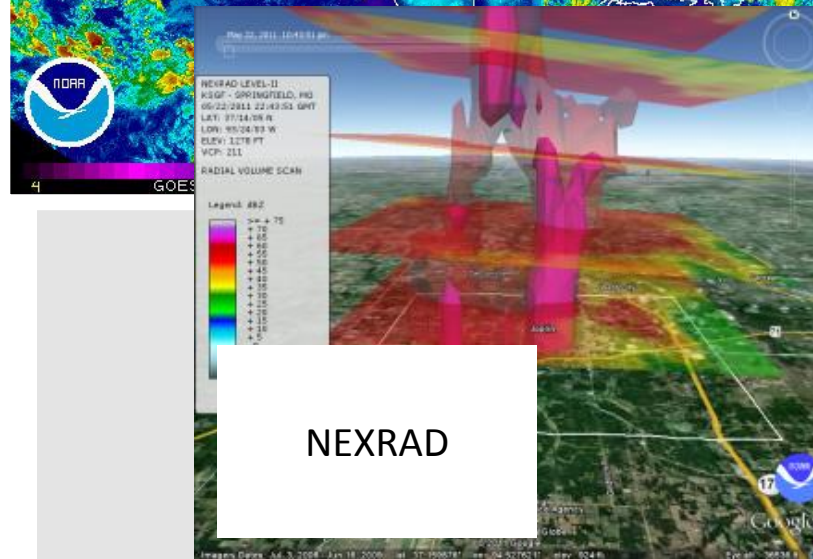
Numerical
Models



30 March 2016

151104/0600V04B GFS 04B-HR TOTAL PCPN (IN)

NEXRAD



CESAS - NAS

Why was NEXRAD Weather Radar selected first?

- Archived NEXRAD data are optimized for *preservation*, not access
- All NEXRAD data are publicly available, but difficult to use
 - unwieldy size (270 TB for compressed Level II alone)
 - specialized format (radial volume scans)
 - resides on NOAA's NCEI tape archive, relatively slow to access
- Highly popular dataset for use in industry
 - Many industry users of realtime and archived NEXRAD data
 - Multiple derivative uses possible - hail, rain, snow, tornados, etc.
- The utilization of entire NEXRAD archive never before realized
- NOAA had recently reprocessed 2001-2012 NEXRAD Level II
 - half of the dataset still resided on disk at the Cooperative Institute for Climate and Satellites in NC (CICS-NC)
 - easy to jump-start initial delivery



NEXRAD #2 in US National Observation Value

Annex I: 2012 EOA Results

This annex provides results for the 145 high-impact observation systems identified from the 362 observation systems assessed by the 13 SBA teams of approximately 300 Federal subject-matter experts. These 145 observation systems are listed in two tiers in the tables below. Impact is indicated with respect to each of the 13 societal themes (12 SBAs and reference measurements), as described in Section 2.2.

Table 1: Tier 1 High-Impact Observation Systems (Ranked Order)

Observation System (Ranked Order)	Agency	Ag&Frst	BioDiv	Climate	Disasters	Ecosys	Energy	HumanHlth	Ocn&Cstl	Space Wx	Trans	WaterRes	Wx	RefMeas
1. Global Positioning System (GPS) satellites	DOD/USAF													
2. Next Generation Weather Radar (NEXRAD)	DOC/NOAA								*					
3. Landsat satellite	DOI/USGS, NASA										*			
4. Geostationary Operational Environmental Satellite System (GOES-NOP)	DOC/NOAA			*		*								
5. National Agriculture Imagery Program (NAIP)	USDA/FSA													
6. Airborne LIDAR	DOC/NOAA, DOD/USACE, DOI/USGS, NSF													
7. Forest Inventory and Analysis (FIA)	USDA/USFS							*						

Impact: *

 * Moderate High Very High Highest

NATIONAL PLAN FOR CIVIL EARTH OBSERVATIONS

PRODUCT OF THE
National Science and Technology Council
Executive Office of the President



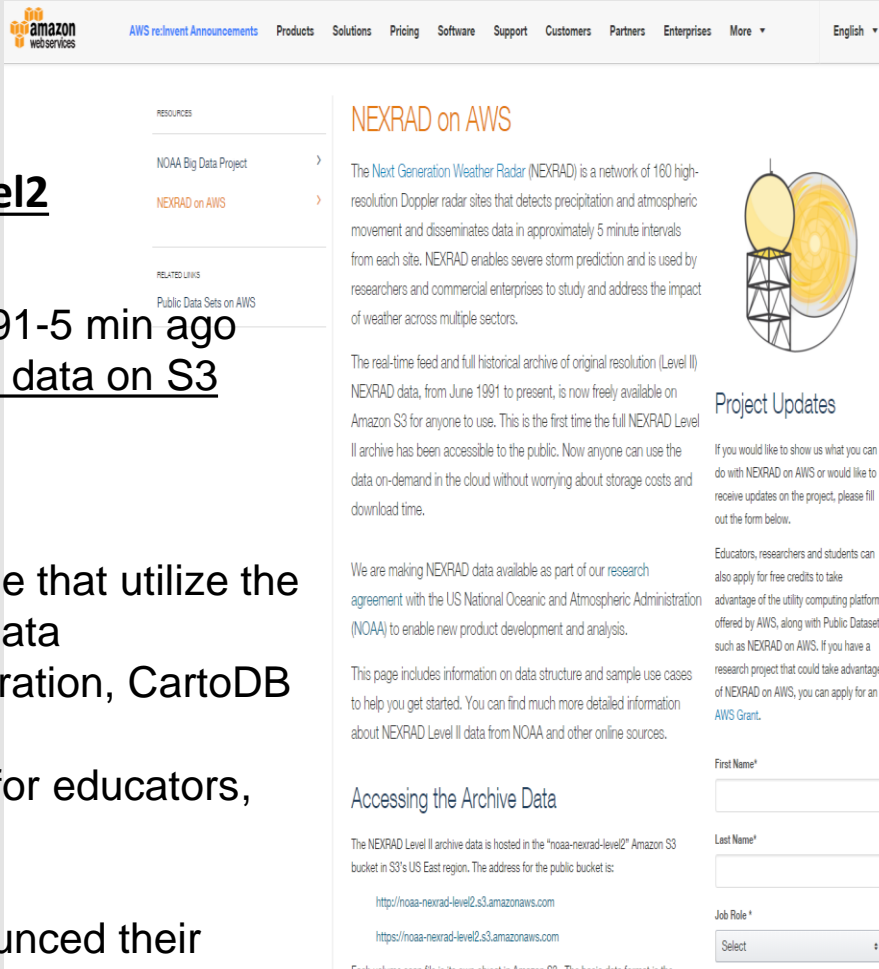
July 2014

National Plan for Civil Earth Observations (2014)

AWS Access to NEXRAD Level II Data

Oct 27, 2015 rollout

- **<https://s3.amazonaws.com/noaa-nexrad-level2>**
- AWS now serving all NEXRAD Level II from 1991-5 min ago
- Single point of access for archived and realtime data on S3
- Free data download available to users
- User-contributed tools and services are available that utilize the AWS platform, e.g. THREDDS installed by Unidata
- Tutorials available from Unidata, Climate Corporation, CartoDB
- Free AWS credits and grants may be available for educators, researchers, and students
- Google, Microsoft, and OCC have not yet announced their access plans for the NEXRAD Level II.



The screenshot shows the AWS website's 'NEXRAD on AWS' page. The header includes the Amazon Web Services logo and navigation links like 'AWS re:Invent Announcements', 'Products', 'Solutions', 'Pricing', 'Software', 'Support', 'Customers', 'Partners', 'Enterprises', and 'More'. A sidebar on the left lists 'RESOURCES' (NOAA Big Data Project, NEXRAD on AWS) and 'RELATED LINKS' (Public Data Sets on AWS). The main content area is titled 'NEXRAD on AWS' and describes the Next Generation Weather Radar (NEXRAD) network. It states that the real-time feed and full historical archive of original resolution (Level II) NEXRAD data, from June 1991 to present, is now freely available on Amazon S3. The page also mentions a research agreement with NOAA and provides links to 'Accessing the Archive Data'. On the right, there is a 'Project Updates' section with a form for users to provide their first name, last name, and job role.

NEXRAD on AWS

The Next Generation Weather Radar (NEXRAD) is a network of 160 high-resolution Doppler radar sites that detects precipitation and atmospheric movement and disseminates data in approximately 5 minute intervals from each site. NEXRAD enables severe storm prediction and is used by researchers and commercial enterprises to study and address the impact of weather across multiple sectors.

The real-time feed and full historical archive of original resolution (Level II) NEXRAD data, from June 1991 to present, is now freely available on Amazon S3 for anyone to use. This is the first time the full NEXRAD Level II archive has been accessible to the public. Now anyone can use the data on-demand in the cloud without worrying about storage costs and download time.

We are making NEXRAD data available as part of our research agreement with the US National Oceanic and Atmospheric Administration (NOAA) to enable new product development and analysis.

This page includes information on data structure and sample use cases to help you get started. You can find much more detailed information about NEXRAD Level II data from NOAA and other online sources.

Accessing the Archive Data

The NEXRAD Level II archive data is hosted in the "noaa-nexrad-level2" Amazon S3 bucket in S3's US East region. The address for the public bucket is:

<http://noaa-nexrad-level2.s3.amazonaws.com>

<https://noaa-nexrad-level2.s3.amazonaws.com>

Each volume scan file is its own object in Amazon S3. The basic data format is the

Project Updates

If you would like to show us what you can do with NEXRAD on AWS or would like to receive updates on the project, please fill out the form below.

Educators, researchers and students can also apply for free credits to take advantage of the utility computing platform offered by AWS, along with Public Datasets such as NEXRAD on AWS. If you have a research project that could take advantage of NEXRAD on AWS, you can apply for an AWS Grant.

First Name*

Last Name*
Job Role *

Select

Steering Users to New Services

- New services are being integrated from within NOAA's NEXRAD sites.
- Archive data ordering system
 - 60% decrease in anticipated data orders from NOAA/NCEI (Feb 2016, S. Ansari)
- Integration into NOAA's free "Weather and Climate Toolkit" application



The screenshot displays the NOAA National Centers for Environmental Information website. The header includes the NOAA logo, the text "NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION", and "NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION". Below this, it states "Formerly the National Climatic Data Center (NCDC)... [more about NCEI](#)". The navigation bar contains links: Home, Climate Information, Data Access, Customer Support, Contact, and About. A search bar is located on the right. The breadcrumb trail reads: Home > Data Access > Radar > Radar Data in the NOAA Big Data Project.

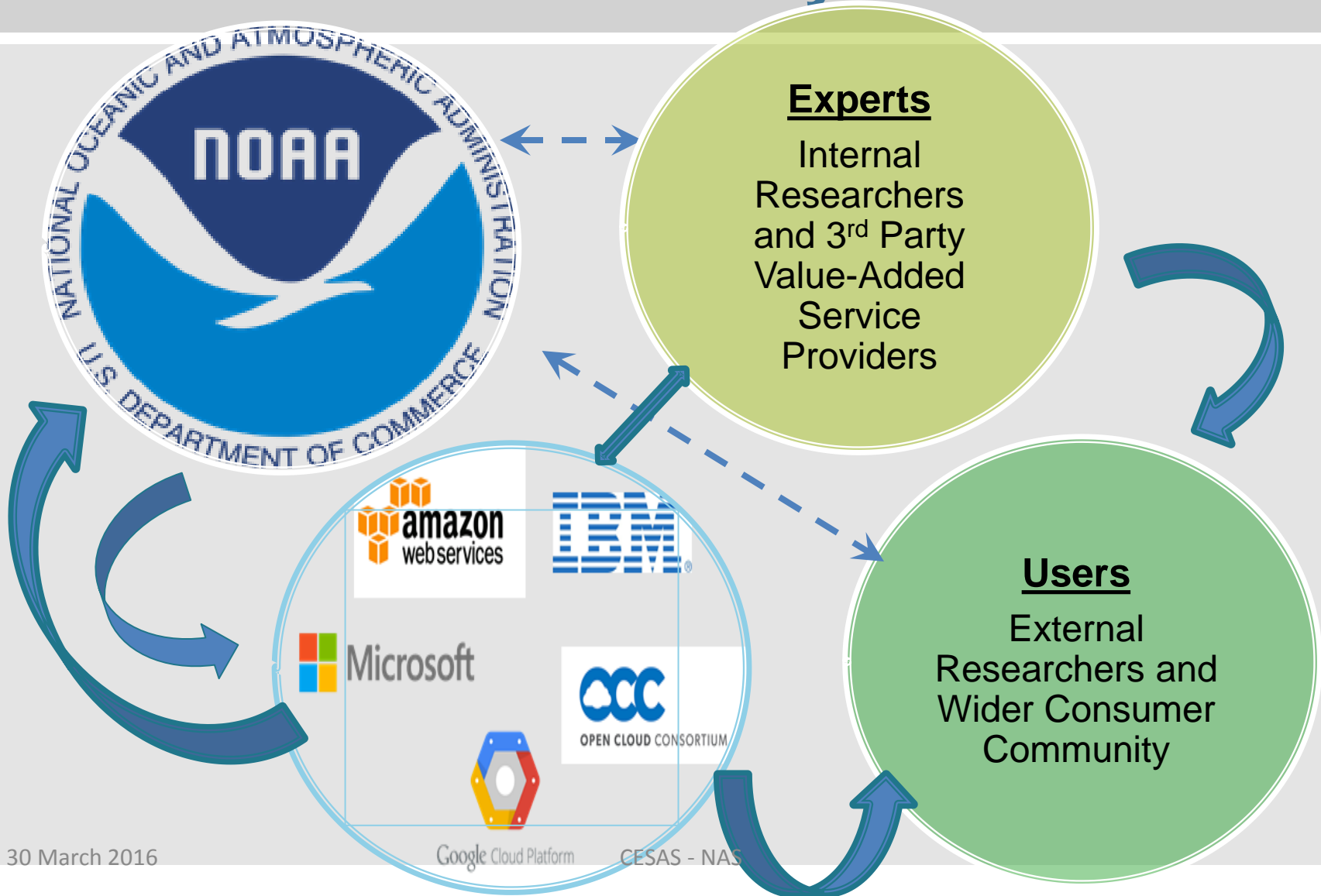
The main content area is titled "Radar Data in the NOAA Big Data Project". It includes a paragraph explaining the NOAA Big Data Project (BDP) as an innovative approach to publishing NOAA's vast data resources and positioning them near cost-efficient high performance computing, analytic, and storage services provided by the private sector. It mentions that this collaboration combines three powerful resources: NOAA's tremendous volume of high quality environmental data and advanced data products, private industry's vast infrastructure and technical capacity, and the American economy's innovation and energy—to create a sustainable, market-driven ecosystem that lowers the cost barrier to data publication. It refers to the [NOAA Big Data Project summary](#) for more information.

Below the paragraph, it states: "Through cooperative activities as a part of NOAA's Big Data Project, NEXRAD data are now freely available through the following cloud infrastructures."

On the left side of the main content area, there is a "Quick Links" section with a list of links: Land-Based Station, Satellite, Radar, Radar Data in the NOAA Big Data Project, Display and Conversion Tools, Decoding Utilities and Examples, Interactive Map Tool, NEXRAD, NEXRAD Radar Products, Terminal Doppler Weather Radar, Terminal Doppler Weather Radar Products, Model, Weather Balloon, Marine / Ocean, Paleoclimatology, Severe Weather, and Blended & Global.

On the right side of the main content area, there is a section titled "Amazon Web Services". It includes the Amazon Web Services logo and a paragraph stating: "The full historical archive of NEXRAD Level-2 data is available for direct download from the Amazon S3 storage or direct access from within the Amazon computing environment." Below this, there are links to "Amazon Documentation", "Amazon Blog", and "NCEI News Release".

The Elements of a Successful BDP Market Ecosystem



So What?

- NOAA's Big Data Partnership is showing that data access and research may be facilitated, at reduced cost, through leveraging effective partnerships with industry.
- Seamless access available to NOAA's holdings across time - users find both historical and realtime NEXRAD Level II data in the same place, in the same way.
 - Historical data provides context for new realtime observations – important for decision-making
- **New business and research opportunities are being created**
- **New applications can be developed FASTER using less bandwidth**
 - data are co-located with the processing capacity
 - Quicken the pace of app development and time-to-market
 - **Faster pace of innovation and scientific discovery**
 - NOAA's NEXRAD recent reprocessing (of 11 years' data) took years
 - Same volume of processing today could now take weeks





Other Lessons Learned in BDP, so far...(1/2)

- Expertise need to be effectively involved throughout the process.
 - “Push” and “Pull” at the same time works the best
 - Sometimes the Collaborators acquire the expertise internally...
- Almost all of NOAA’s open data are available, digitally, today...if you know who to ask, or where to look.
- True Big Data applications now limited by their business cases
 - Single valuable datasets or small collections, versus truly diverse and large datasets
 - Without easy access to Big Data, how do you know what it will bring? How to decide to commit resources within BDP?



Other Lessons Learned in BDP, so far...(2/2)

- Easy access to archived data is valuable for dynamic provision of the historical/statistical context for new data or realtime data flows.
- Traditional data-centered economic value models are not necessarily easily adaptable to the Big Data methodology
- There may be more social problems than technical problems.
- How can NOAA effectively steward data held by others?



Other Opportunities Ahead

- The rise of the Platform
 - Platform = data + processors + storage + apps
 - Scientists used to go to the data. Then the data came to them. Now they all can coexist in a shared, virtual, computing environment.
 - Has Platform utilization become “simply” an acquisitions and governance problem? Or do social problems need to be addressed first?
- The US Digital Service and its related entities (e.g. the Commerce Data Service, GSA 18F)
 - Focusing on enabling data discovery and use
 - E.g. the Commerce Data Usability Project



Summary

- The full utilization of Earth Science data is a Big Data challenge since it requires the provision of adequate expertise, involves many datasets, and is challenging conventional value models.
- “Big Data” success requires not just access to the data, but the expertise (algorithms, workflows, interpretive skill) as well.
- Government and Industry partners may be able to facilitate research activities by leveraging the inherent value in Earth Science data.
- Through the BDP CRADA, NOAA has moved the NEXRAD Level II dataset first, based on market need and data opportunity. AWS experience Oct 2015 appears to be heading towards success so far...
 - New applications created, increased use of AWS infrastructure, reduced load on NOAA systems
- NOAA and the Collaborators have not really gotten deep into Big Data territory yet, so far mainly enabling wider use of unwieldy, hard to access and high volume datasets.



Acknowledgements

Many thanks to:

- NOAA: Amy Gaskins, Alan Steremberg, Maia Hansen, Steve Ansari, Steve Del Greco, Jeff de la Beaujardiere, Brian Nelson, Tony LaVoi, Jay Morris, Carlos Rivero, Ken Casey, Ken Knapp
- NC State University / CICS-NC: Otis Brown, Jonathon Brannock, Lou Vazquez, Scott Stevens

NOAA's Big Data Collaborators and their partners involved in the NEXRAD project

- Amazon: Arial Gold, Jed Sundwall, Jeff Layton
- Climate Corporation: Adam Pasch, Valliappa Lakshmanan,
- Unidata: Jeff Weber
- Microsoft: Sam Khoury, Sid Krishna
- Google: Eli Bixby, Tino Tereshko, Amy Unruh, Tanya Shastri, Ossama Alami
- Open Commons Consortium: Maria Patterson, Walt Wells

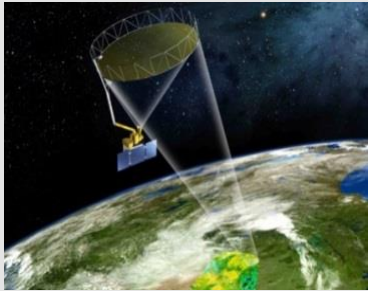
Thank You - Discussion

Questions?

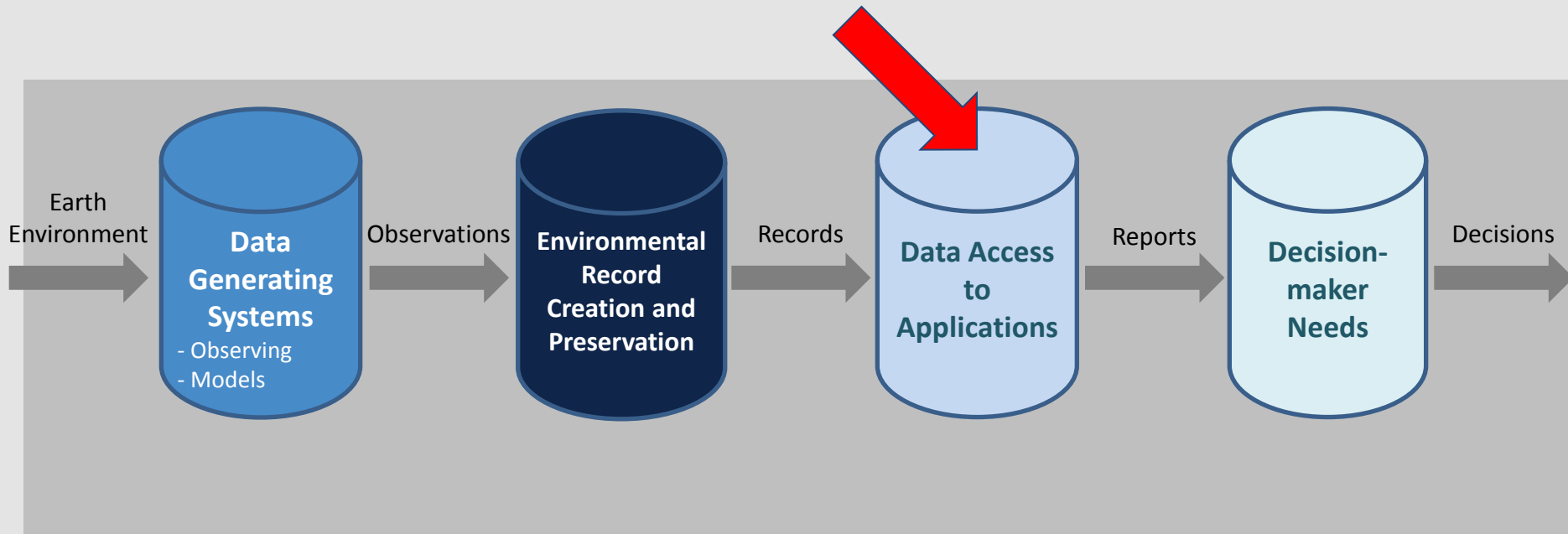




Impact on the Information Lifecycle



- NOAA makes foundational investments in environmental information production and preservation
- BDP promotes increased data access and encourages application development across private and public sectors.



NEXRAD Data Transfer

- Archived Level II data from NESDIS' National Centers for Environmental Information (NCEI)
 - Over 270 TB for compressed Level II volume scan file (>1 PB uncompressed)
 - Move from the NCEI tape archive to disks at CICS-NC
 - CICS-NC acting as middleman allowed minimal impact to NOAA's data operations
 - Updates from NCEI archive operations as new data are pushed to the archive
- Realtime Level II data from NOAA's National Weather Service
 - Established through Unidata, part of the Open Commons Consortium (OCC)
- Data moved to 4 of the 5 Big Data Partnership Collaborators
 - Amazon Web Services (entire archive plus realtime)
 - Microsoft (entire archive)
 - Google (entire archive)
 - Open Commons Consortium (2015 plus realtime)





Big Data Partnership “Rules”


- **Level Playing Field, for all interested companies via Collaborators**
 - All NOAA Data, available equally, with no privileged access
 - NOAA-sourced data are “free and open” - only cost recovery allowed
- **5 individual CRADAs with NOAA, 3 year terms + 2 one-year options**
- **CRADAs can be ended early with appropriate notice**
- **No funds may move from NOAA to Collaborators**



Big Data Partnership Methodology

- **Collaborators & their partners identify datasets of interest**
- **Demonstrate business use case(s), with no restrictions on open distribution of data**
- **Develop a strategy for data sharing/delivery from NOAA to BDP Collaborator(s)**
- **Engage NOAA subject matter experts, BDP Collaborators, and their Associates/Partners for technical interchanges**
- **Collaborators and their Partners create applications**
- **NOAA continues all of its existing data services**
 - **No interruption of services to customers**
 - **BDP activities are an augmentation of existing services**

Steering Users to New Services

 **NOAA** NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

[Home](#) [Contact Us](#) [About NCEI](#) [Help](#)

NCEI > Radar Data > NEXRAD Inventory > Choose Date / Product

NEXRAD Archive Access

Data Access
Home
Select By Map
Select By List
Select By County, City, Zip Code (Climate Data Online)
Historical Reflectivity and Coverage Maps
Select By Archive File (Bulk Order)

Documentation
Archive and Access Statistics
Overview, History
NEXRAD Product List
Network Metadata
Radar Operations Center

External Resources
NOAA/NWS Current Radar
NOAA Training (WDTB)

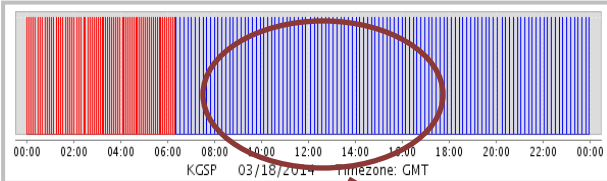
NEXRAD Inventory: View Results and Order Data

KGSP - GREER, SC
[\(Site Metadata\)](#)

Period of Record:
Level-II: 07/14/1995 to 03/19/2014

[03 / 17 / 2014](#) **03 / 18 / 2014** [03 / 19 / 2014](#)

Level-II Base Data



00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00 00:00
KGSP 03/18/2014 Timezone: GMT

— Clear Air Mode — Precip Mode — Maintenance Mode — Unknown Mode

Enter Email Address:

Start Time: 00:00 GMT End Time: 24:00 GMT

[View Actual Timestamps / Op. Mode / VCP](#)

NEXRAD Data - Direct Download

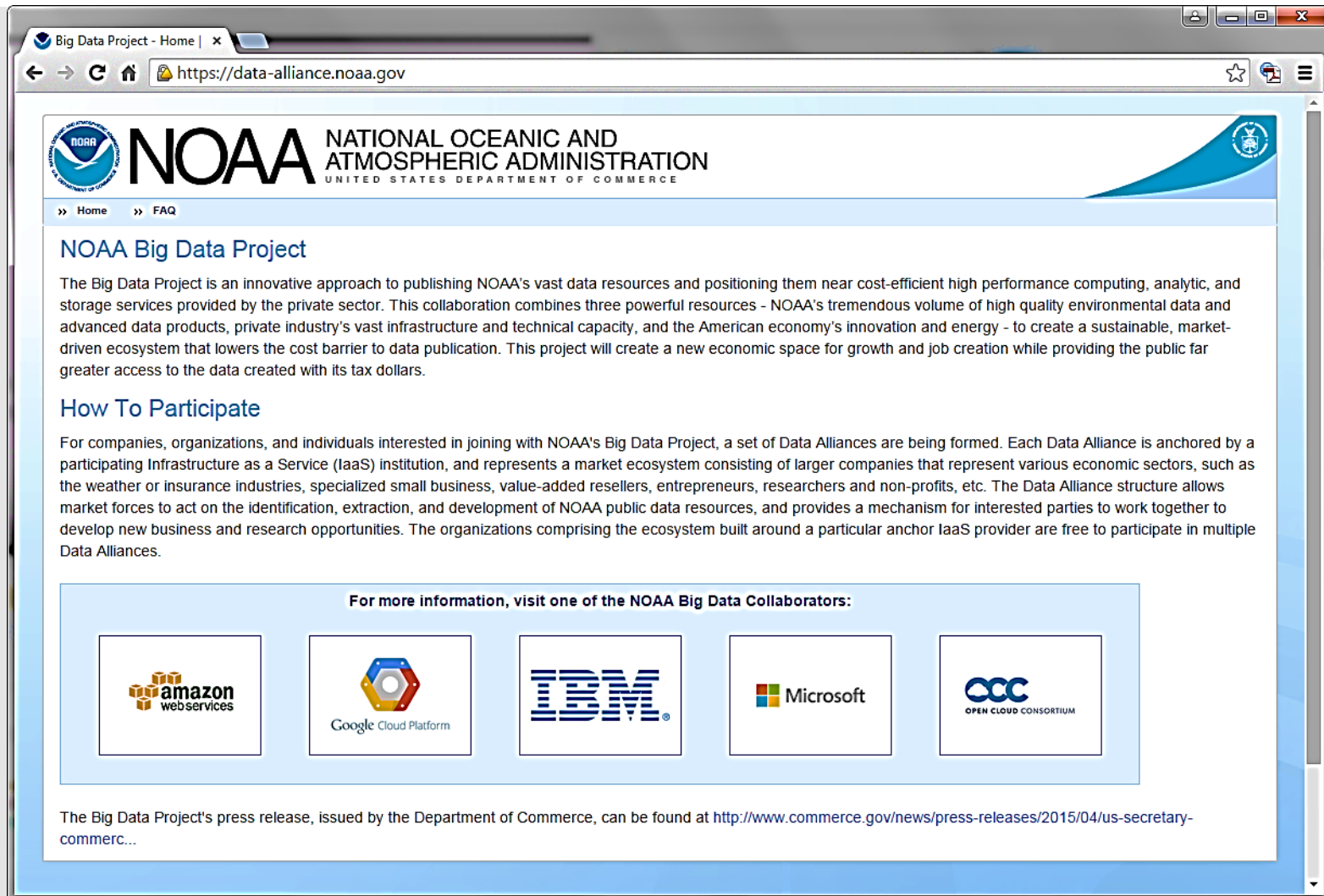
Other BDP Collaborators' sites will be added as they roll out their data services



Role of NOAA in the BDP Market Ecosystem

- NOAA's data are widely available, with “free and open” access
- NOAA's role will be to provide objective scientific expertise, ensure long-term preservation and sound data management (scientific data stewardship)
- Data + Expertise + Need = Opportunity


<https://data-alliance.noaa.gov>



The screenshot shows a web browser window with the address bar displaying <https://data-alliance.noaa.gov>. The page header features the NOAA logo and the text "NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION" and "UNITED STATES DEPARTMENT OF COMMERCE". Below the header, there are navigation links for "Home" and "FAQ". The main content area is titled "NOAA Big Data Project" and contains a paragraph describing the project's goals and a section titled "How To Participate" which explains the Data Alliance structure. At the bottom, there is a section titled "For more information, visit one of the NOAA Big Data Collaborators:" which lists five collaborators: Amazon Web Services, Google Cloud Platform, IBM, Microsoft, and the Open Cloud Consortium (OCC).

Big Data Project - Home | x

https://data-alliance.noaa.gov

 **NOAA** NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
UNITED STATES DEPARTMENT OF COMMERCE

» Home » FAQ






NOAA Big Data Project

The Big Data Project is an innovative approach to publishing NOAA's vast data resources and positioning them near cost-efficient high performance computing, analytic, and storage services provided by the private sector. This collaboration combines three powerful resources - NOAA's tremendous volume of high quality environmental data and advanced data products, private industry's vast infrastructure and technical capacity, and the American economy's innovation and energy - to create a sustainable, market-driven ecosystem that lowers the cost barrier to data publication. This project will create a new economic space for growth and job creation while providing the public far greater access to the data created with its tax dollars.

How To Participate

For companies, organizations, and individuals interested in joining with NOAA's Big Data Project, a set of Data Alliances are being formed. Each Data Alliance is anchored by a participating Infrastructure as a Service (IaaS) institution, and represents a market ecosystem consisting of larger companies that represent various economic sectors, such as the weather or insurance industries, specialized small business, value-added resellers, entrepreneurs, researchers and non-profits, etc. The Data Alliance structure allows market forces to act on the identification, extraction, and development of NOAA public data resources, and provides a mechanism for interested parties to work together to develop new business and research opportunities. The organizations comprising the ecosystem built around a particular anchor IaaS provider are free to participate in multiple Data Alliances.

For more information, visit one of the NOAA Big Data Collaborators:



The Big Data Project's press release, issued by the Department of Commerce, can be found at <http://www.commerce.gov/news/press-releases/2015/04/us-secretary-commerce...>