Informing NOAA's Mission with the Big Data Project experience

Dr. Edward J. Kearns
National Oceanic and Atmospheric Administration

Workshop on Data and Computational Science Technologies for Earth Science Research

Committee on Earth Science and Applications from Space

National Academies of Science 4 October 2016

Outline

- Update on the NOAA Big Data Project (BDP)
  - *research activity with NOAA and its Collaborators*
  - *recent statistics from the first mature data project*

- What we are learning from the BDP experience
  - *"Portals versus Platforms"*
  - *data stewardship roles and responsibilities*

- Some thoughts on future impacts of advances in information and communication technologies

NOAA's Mission: Science, Service and Stewardship

1. To understand and predict changes in climate, weather, oceans and coasts

2. To share that knowledge and information with others

3. To conserve and manage coastal and marine ecosystems and resources

Big Data Project Recap

NOAA's Big Data Project (CRADA) began in April '15

Leverage the value of NOAA's data to increase their utilization?

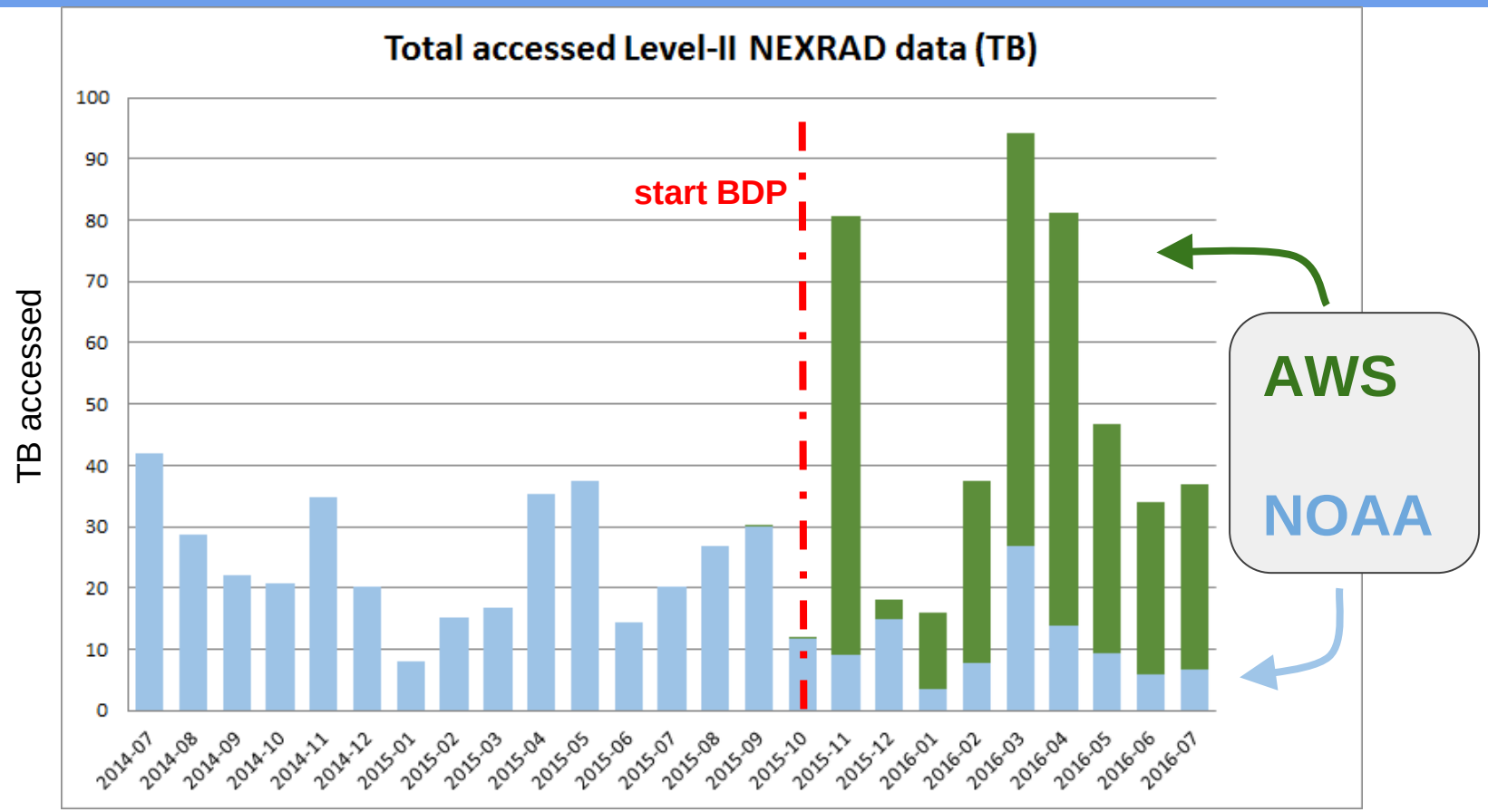
Keys:

- NOAA's data
- Industry's infrastructure expertise
- NOAA's subject matter expertise
- Level playing field

OPEN COMMONS CONSORTIUM

NEXRAD Weather Radar Data

2015: Moved 850 TB to NOAA BDP Collaborators



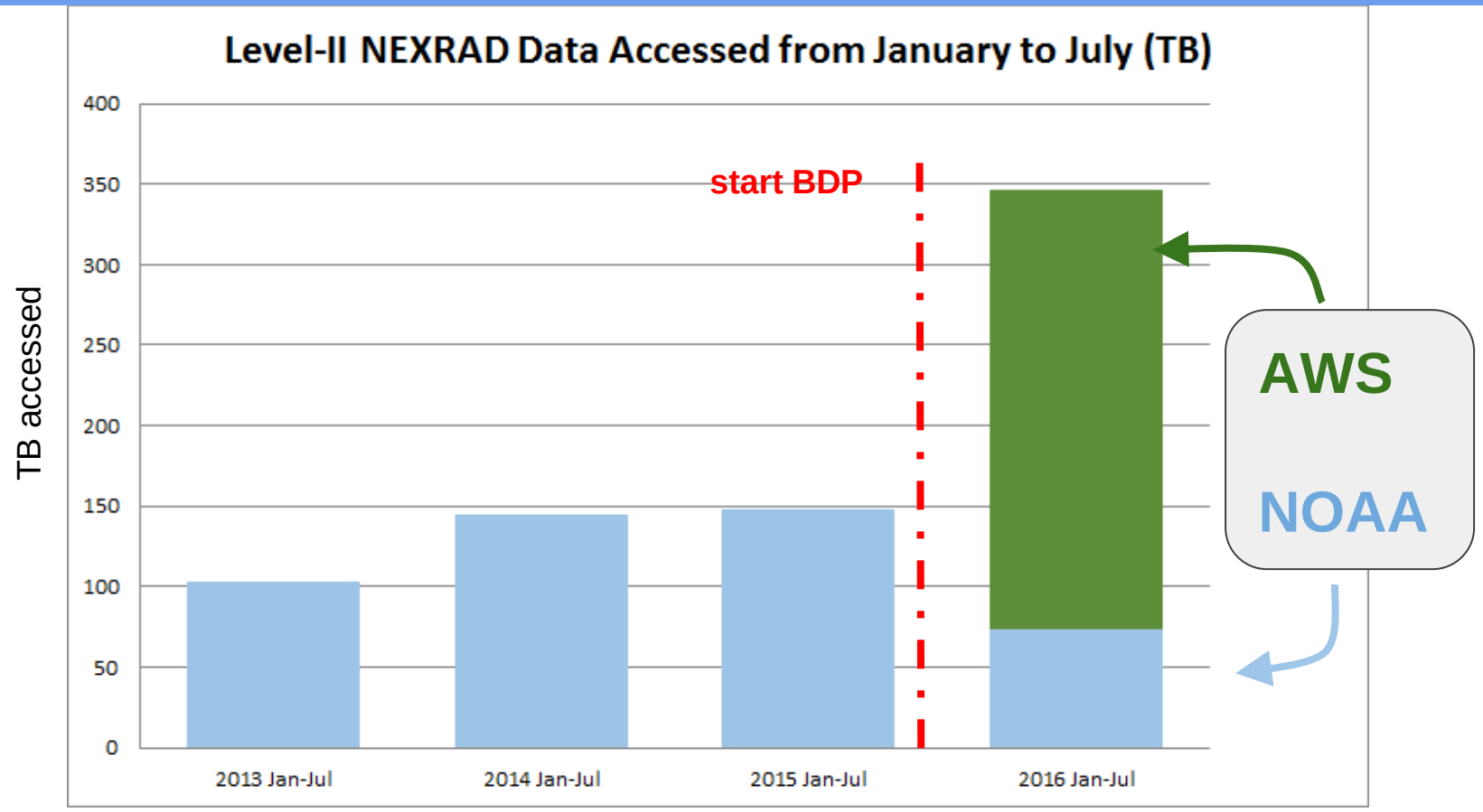
AWS: Oct '15 https://s3.amazonaws.com/noaa-nexrad-level2 **(1991+)**

OCC: Jun '16 http://occ-data.org/NOAANEXRAD/ **(2015+)**

(S. Ansari et al, 2016)

NEXRAD Weather Radar Data

2015: Moved 850 TB to NOAA BDP Collaborators



(S. Ansari et al, 2016)

NEXRAD on AWS Statistics

- 80% of NEXRAD archive orders are now fulfilled by AWS
- Single access point for both archived and realtime data
- 64% of the NEXRAD data stayed on the AWS platform

Data utilization has increased by 2.3 times at AWS, at no net cost to the US taxpayer

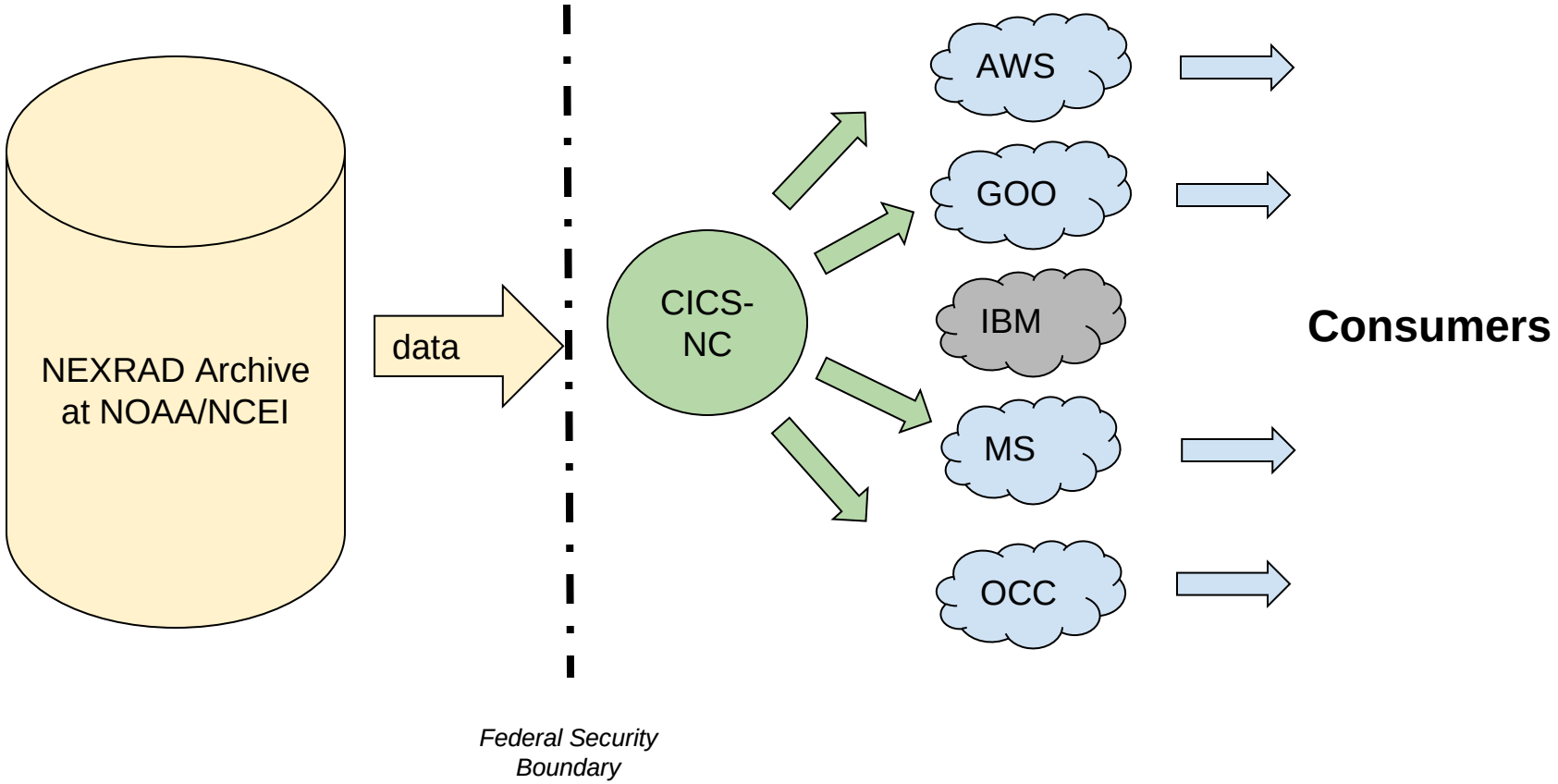
- Faster: jobs that took 3+ years now take only a few days
- Cheaper: loads on NOAA archives are down over 50%

What must we determine though the BDP during the CRADA?

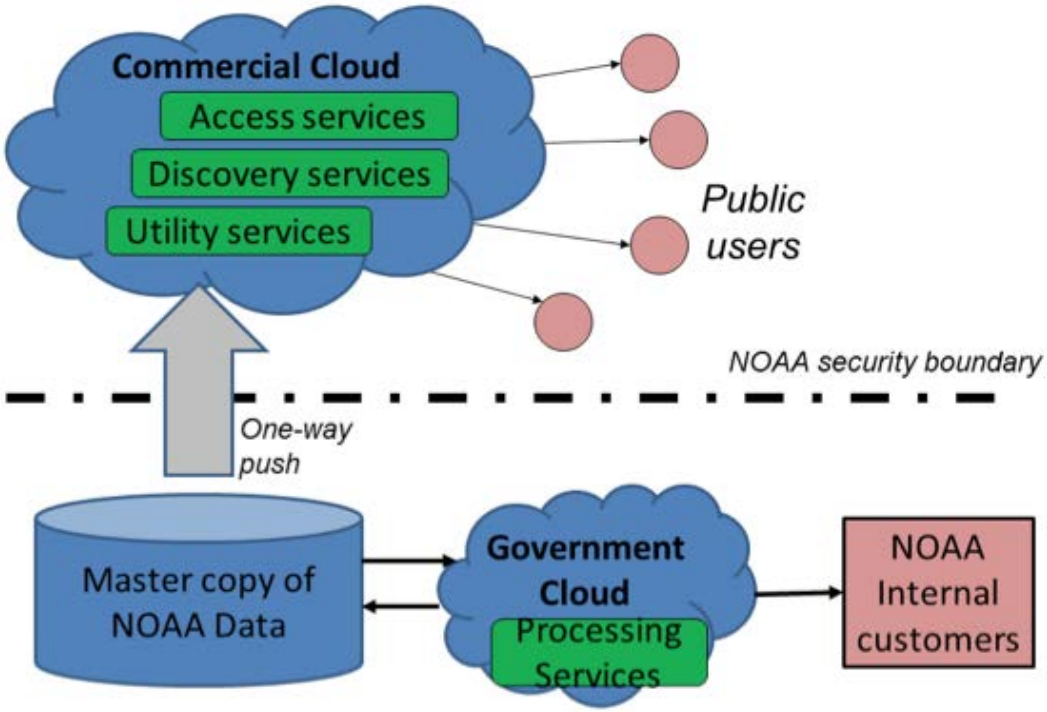
- Optimal business relationship between NOAA and its Collaborators
  - Cost recovery strategies?
  - Scalability across NOAA data holdings?
  - Value of data, Value of stewardship, Value of services?

- Roles and responsibilities regarding data quality
  - Goes well beyond legal liability issues
  - Ensure quality is understood throughout the value chain

“From One, To Many” Model (NEXRAD BDP Example)



NOAA EDM Framework



NOAA Environmental Data Management Framework (2013)

Appendix C, Fig 8

https://nosc.noaa.gov/EDMC/framework.php

Portals versus Platforms

- Are there significant advantages for Understanding and Sharing in having NOAA's data on a Platform (processing, storage, software) instead of a Portal?
  - Faster utilization, Modern tools, New applications?
    - Artificial Intelligence and Deep Learning
  - Community approach, open/transparent
  - Easier reproduction/verification of research results?
    - Processing on the fly
  - Dual use - both for Internal NOAA Use and Public Dissemination?

- How will these Platforms be maintained?
  - Should NOAA populate and steward the data on the platform?
  - Include Software/code for analysis and utilization?
  - Single active copy (store) of data?
    - Serves both Sharing with others and Research by NOAA

Data Quality

- Does NOAA “own” the data quality issue?
- NOAA can verify data up to the hand-off point today
  - usually via checksums or digital signatures
- **How to enable verification and/or provenance after hand-off, when the data formats have been changed when incorporated into modern tools?**
  - technical solutions - registries and digital signatures

- Does it make sense to have NOAA steward the data on the commercial cloud?
  - a labor-intensive and expensive activity
  - NOAA owns the expertise
  - What is the added value of a expertly-curated dataset?

Other Issues

- What is meant by “preservation” in the cloud?
  - Is the “gold copy” or “master copy” concept still valid?
  - Verification and authentication services
  - Focus on data management instead of on hardware/storage
- Intellectual Property rights must be identified early
  - joint development of products and services
  - can registries also facilitate proper compensation?
- How will the emerging role of the “Chief Data Officer” across both government and industry enable new partnerships to utilize data?

Questions and Discussion

Acknowledgements

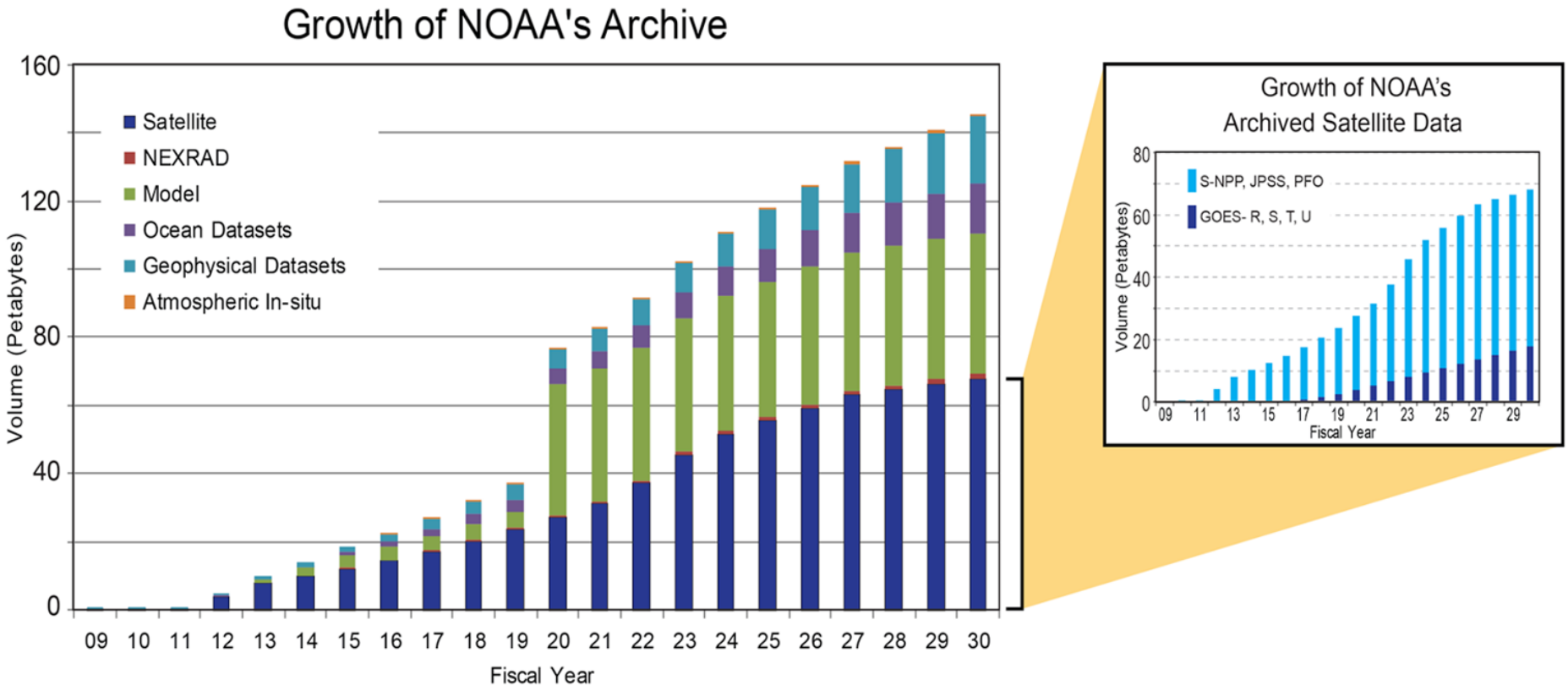
Many thanks to:

- NOAA: Brian Eiler, Zach Goldstein, Dave Michaud, Glen Talia, 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
- NC State University / CICS-NC: Otis Brown, Scott Wilkins, Jon Brannock, Lou Vazquez, Scott Stevens, Paula Hennon, Andrew Buddenberg

NOAA's Big Data Collaborators and their partners (not an all inclusive list)

- Amazon: Jed Sundwall, Arial Gold (now @DOT), Jeff Layton
- Microsoft: Sam Khoury, Sid Krishna
- Google: Will Curran, Matt Hancher, Eli Bixby, Tino Tereshko, Amy Unruh, Tanya Shastri, Ossama Alami, Valliappa "Lak" Lakshmanan (formerly @Climate)
- Open Commons Consortium: Walt Wells, Maria Patterson, Zac Flamig
- Unidata: Mohan Ramamurthy, Jeff Weber
- IBM: James Stevenson, Stefani Jones, Mary Glackin, Peter Neilley
- The Climate Corporation: Adam Pasch

Archive Projections for NOAA data



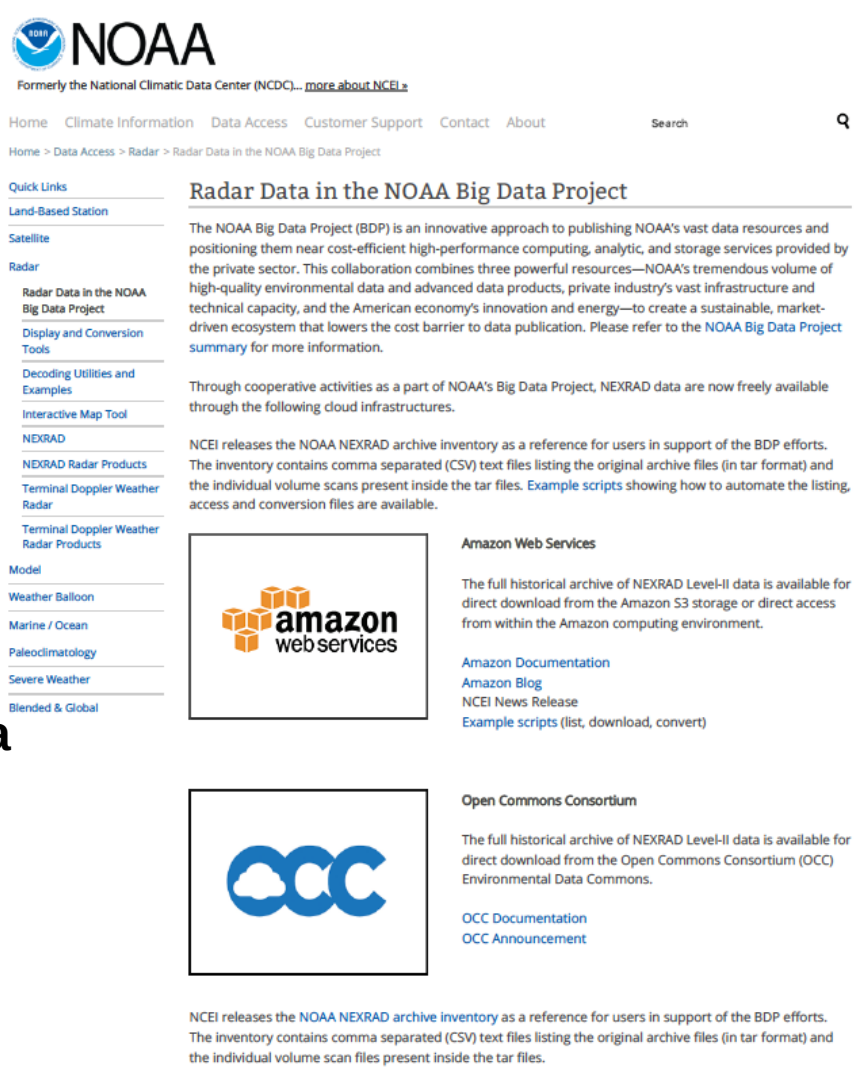
Courtesy Steve Del Greco/Ken Casey, NOAA/ NCEI

Steering Users to New Services

From NOAA's web site and tools (Weather and Climate Toolkit), offer users the choice of using data on Collaborators' cloud services.

NOAA also provides data inventories:

http://www1.ncdc.noaa.gov/pub/data/radar/archive-inventory/



NOAA

Formerly the National Climatic Data Center (NCDC)... more about NCEI »

Home Climate Information Data Access Customer Support Contact About Search

Home > Data Access > Radar > Radar Data in the NOAA Big Data Project

Quick 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
Blended & Global

Radar Data in the NOAA Big Data Project

The NOAA Big Data Project (BDP) 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. Please refer to the NOAA Big Data Project summary for more information.

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

NCEI releases the NOAA NEXRAD archive inventory as a reference for users in support of the BDP efforts. The inventory contains comma separated (CSV) text files listing the original archive files (in tar format) and the individual volume scans present inside the tar files. Example scripts showing how to automate the listing, access and conversion files are available.

amazon web services

Amazon Web Services

The full historical archive of NEXRAD Level-II data is available for direct download from the Amazon S3 storage or direct access from within the Amazon computing environment.

Amazon Documentation
Amazon Blog
NCEI News Release
Example scripts (list, download, convert)

OCC

Open Commons Consortium

The full historical archive of NEXRAD Level-II data is available for direct download from the Open Commons Consortium (OCC) Environmental Data Commons.

OCC Documentation
OCC Announcement

NCEI releases the NOAA NEXRAD archive inventory as a reference for users in support of the BDP efforts. The inventory contains comma separated (CSV) text files listing the original archive files (in tar format) and the individual volume scan files present inside the tar files.