



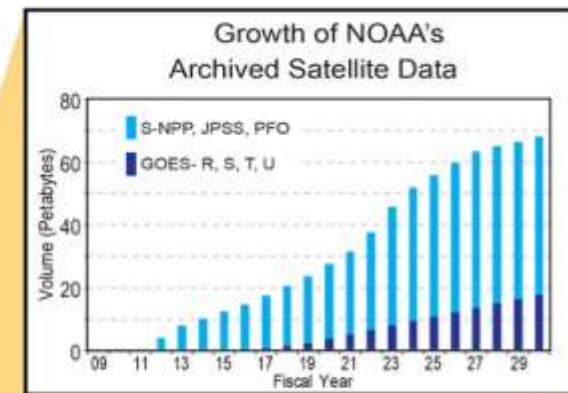
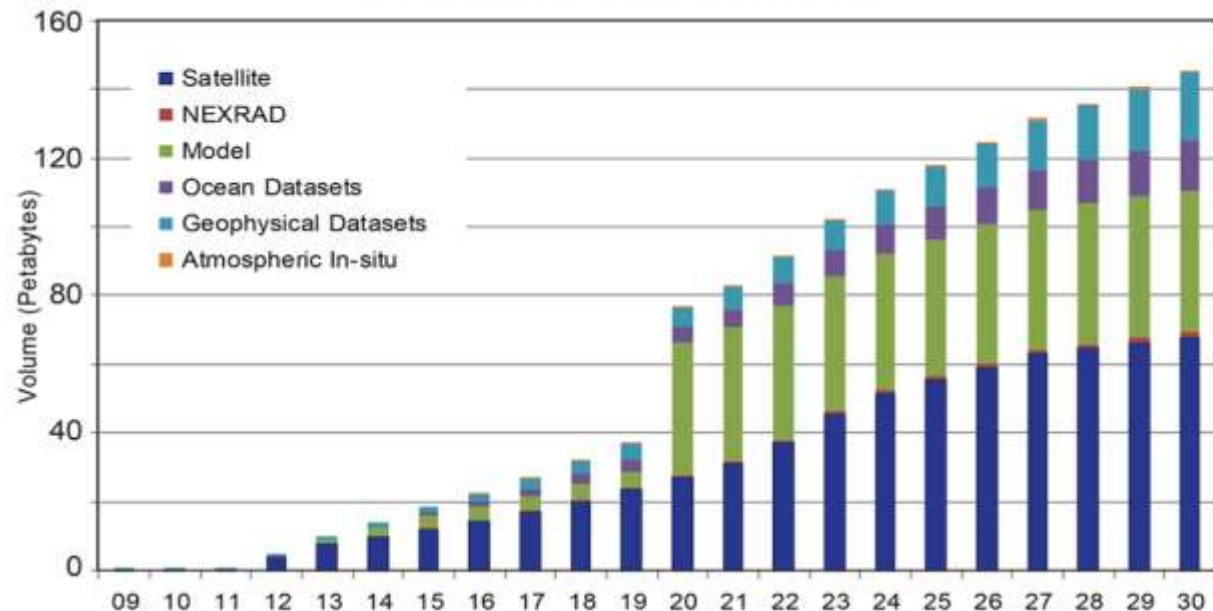
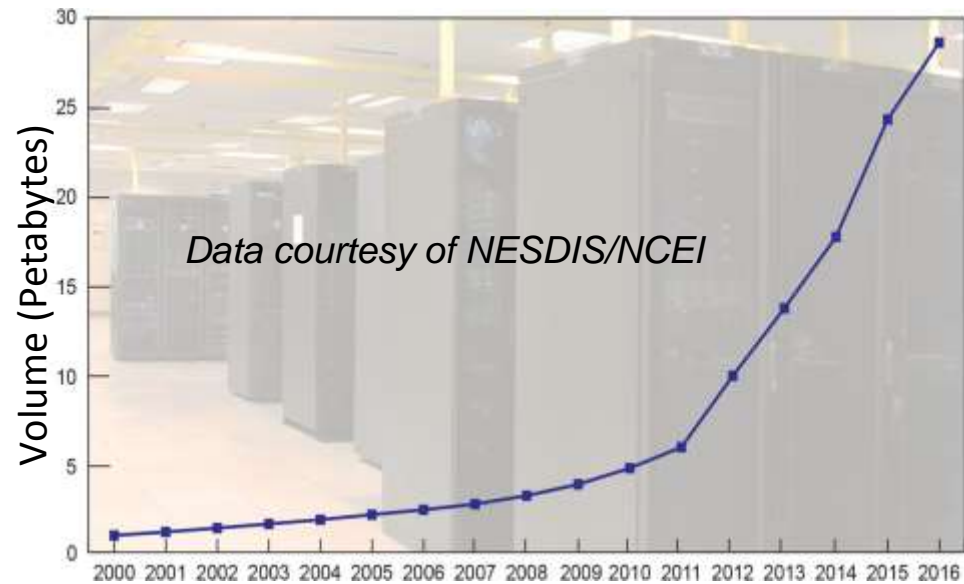
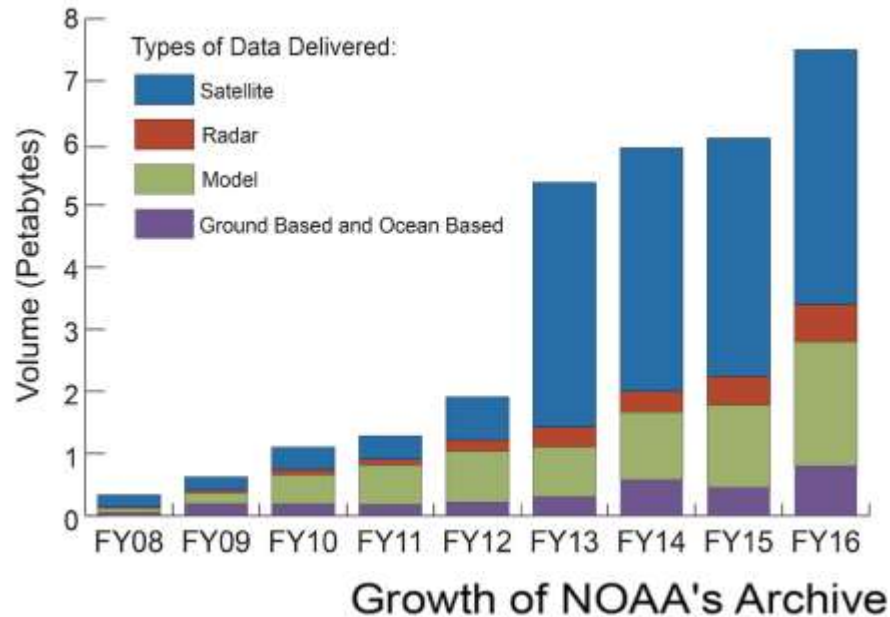
# Big Data and NOAA Science

Dr. Edward Kearns  
NOAA Chief Data Officer  
National Academies of Science  
Space Studies Board, Irvine CA  
*Nov 2, 2017*

# Why is NOAA Interested in Big Data?

- The **volume and velocity** of NOAA observations and model outputs continues to rise exponentially.
- NOAA's **science-based products and services** can benefit from Big Data approaches.
- NOAA data are **increasingly popular and valuable**
  - NOAA struggles to keep up with **demand** for its data.
- Enable new **economic and research opportunities**

# NOAA/NCEI's Environmental Data Archive



# Some Possible Services and Activities with Big Data implications

**Fisheries:** Fish catch monitoring, stock and habitat assessment, genomics research, IUU Fishing

**Oceans and Coasts:** sea level rise, inundation estimates, and disaster response

**Weather and Climate:** Forecaster Assistance, Model output utilization, V&V, Data assimilation, Multi-model Ensembles, Combining deterministic and probabilistic approaches

**Satellite and Archive:** Data exploitation, Product retrievals, Cal/Val, Data Quality Assurance, AI/ML training

**Operations:** aircraft/ship logistics, glider/buoy/drone surveys



# Use Big Data, AI/ML techniques to aid humans in the delivery of services

Augment Forecasters' skills by providing Big Data and ML/AI applications to help them execute NOAA's Mission

Many data sources and inputs, in a high-stakes environment: satellite, radar, aircraft, balloons, models, meso-nets, IOT, social media, etc.



# Example of Enhanced Data Presentation for Services: AWIPS-II and ProbSevere

Model output are shapefiles contoured around radar storm cells.

Enhancement designed for overlay atop radar reflectivity—but can be overlaid on any field (satellite, radar velocity, etc.).

Sampling offers readout of model probability as well as each model predictor.



SVR PROB: 97%  
- Env MLCAPE: 3296 J/kg  
- Env EBShear: 55.5 kts  
- NWS MESH: 2204Z 1.06 in.  
- Norm Vert Growth Rate (Max): 2130Z 5.60 %/min (strong)  
- Glaciation Rate (Max): 2125Z 0.07 /min (moderate)  
- Object ID: 26589  
- Flash Rate: 2204Z 32.2 ft/min  
Lightning Jump: 8.3 sigma  
57.1dBZ

# NOAA Big Data Project

Evaluate partnership opportunities with Cloud Computing industry

- Provide cloud-based access to NOAA's open data
  - Copies of NOAA data accessed from Partners' systems
  - Improved federal cybersecurity posture
- Cost Avoidance for public data access
  - Most popular datasets bring largest burden on NOAA systems
- Better Level of Service to customers
  - Users can utilize data faster without downloading
- **This is not *just* about open data access**
  - **Can accelerate data utilization...**
  - **...and thus improve societal impacts, research and business opportunities**

# BDP Basics



- Cooperative Research and Development Agreements
  - 5 separate but identical 3-year agreements
- Industry provides access to NOAA's open data to all
  - Data remain open, are not to be sold
  - Collaborators monetize services based on data
  - Dropped typical egress charges
  - NOAA provides data and expertise
- Combines 3 powerful resources based on NOAA's open data:
  1. NOAA's science and subject matter expertise
  2. Industry's data storage and access expertise
  3. Cloud's scalable and on-demand processing capability



**NOAA**  
Data Expertise

**CRADA Collaborators**  
Infrastructure Expertise



**End User**  
Wider Consumer Community

**Third Party Partner**  
Value-Added Services

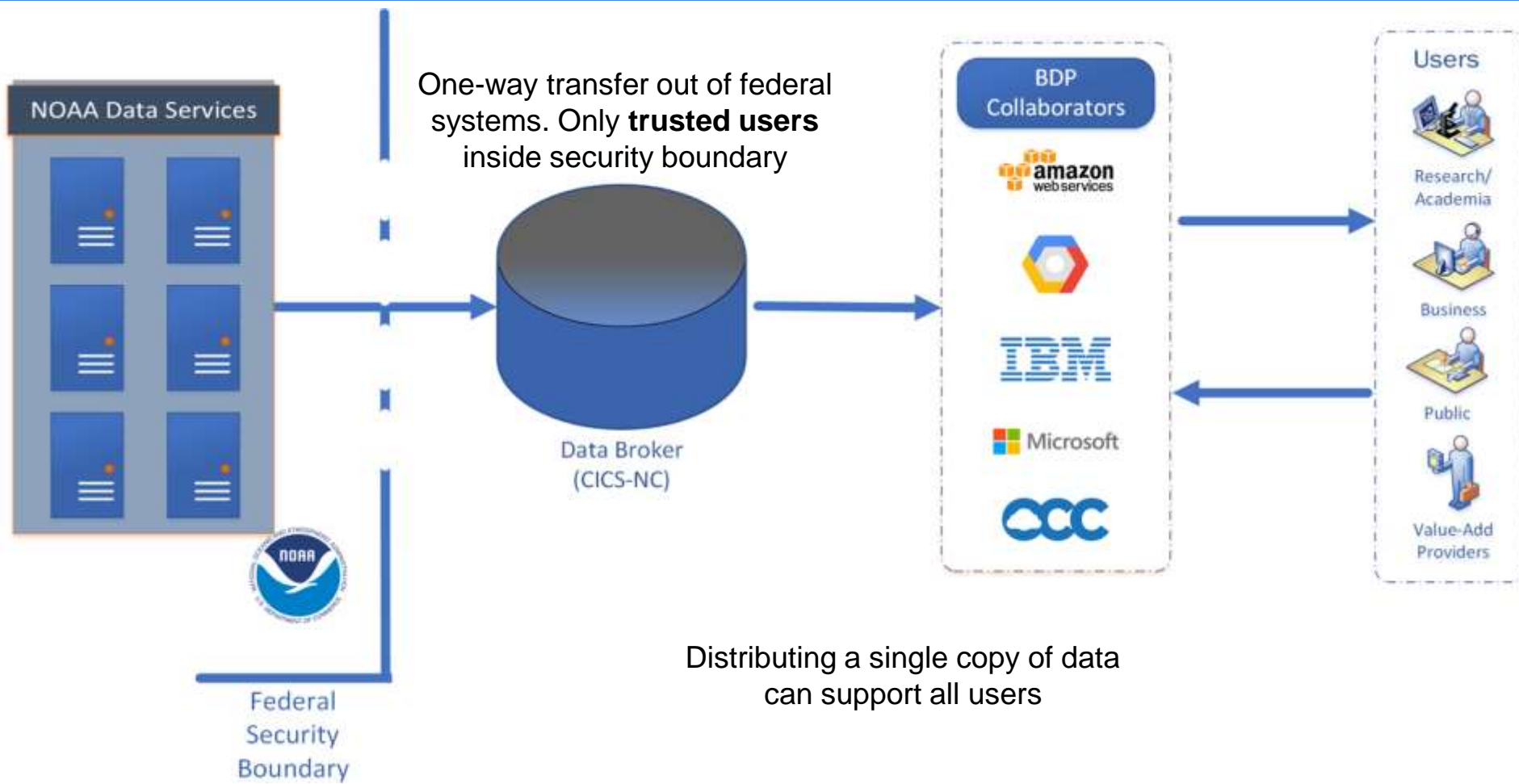
# Two different examples of how to increase data utilization

- Weather Radar: NEXRAD on AWS (270 TB)
  - 2.3x increase in usage over past
  - Redirected orders from NOAA to AWS as option for users
  - *Ansari et al, BAMS 2017*
- Climate & Weather: GHCN-M, GSOD on Google (< 0.1 GB)
  - 800,000 data requests between Jan and Apr 2017
  - **1.2 PBs of data delivered**
  - **100x or more increase in usage over the past**
  - No redirects - completely organic utilization by users of the tools
  - Integration of data into existing tools more effective

**Who is going to do the integration?**

# Data Broker Role

One-to-Many, Limited by Cloud Infrastructure Only



# GOES-16 BDP Demo Live as of July 12, 2017: Initial Distribution Statistics

Cooperative Institute for Climate and Satellites - North Carolina (CICS-NC) is helping NOAA by providing feeds of the GOES-16 data from the NOAA Ground System (as an authorized user) to the BDP CRADA Collaborators.

- BDP is offering 5 validated feeds to the CRADA Collaborators
  - Timing - as fast as they appear at NOAA distribution point
  - Single bounce of data through CICS-NC systems, w/checksums
  - Minimizes load on NOAA's operational systems and networks
- Observed additional latencies from CICS-NC transfer mechanism
  - From NOAA Ground System to BDP Collaborator platforms
  - Maximum additional latency: 2 to 3 min (full disk ABI, Band 2)
  - **Typical Range of additional latency: 30 sec - 3 min**

# Big Data Project

## Collaborators' Data Offerings

- **AWS**
  - <https://aws.amazon.com/noaa-big-data/>
- **Google Cloud Platform**
  - <https://cloud.google.com/bigquery/public-data/> *(see NOAA listings on left)*
- **IBM**
  - <https://noaa-crada.mybluemix.net/>
- **Microsoft**
  - No public services to date
- **Open Commons Consortium**
  - <http://edc.occ-data.org/>



# Big Data Project and Open Data Challenges

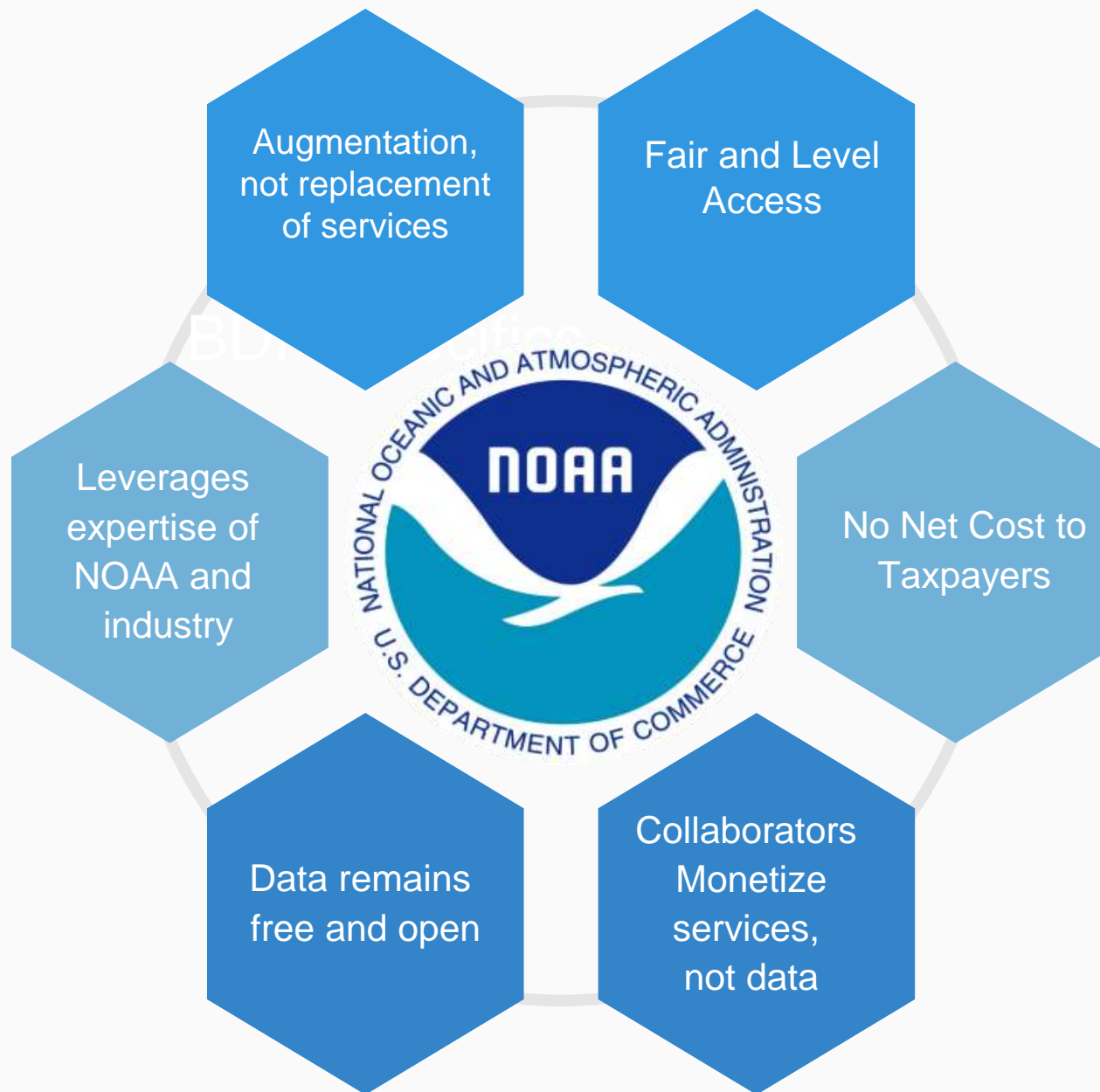
- How can NOAA best adopt modern Big Data tools?
- How well do we understand the Big Data market?
  - All NOAA's data commercially-viable in this model?
  - The role of researchers in the ecosystem?
- How to sustain data-centric public—private partnerships?
- How to best steward numerous large, complex datasets?
  - Extend NOAA's "brand" on widely distributed data?
  - How to ensure data authenticity at low cost?

# Discussion

[ed.kearns@noaa.gov](mailto:ed.kearns@noaa.gov)

#NOAABigData

<http://www.noaa.gov/big-data-project>



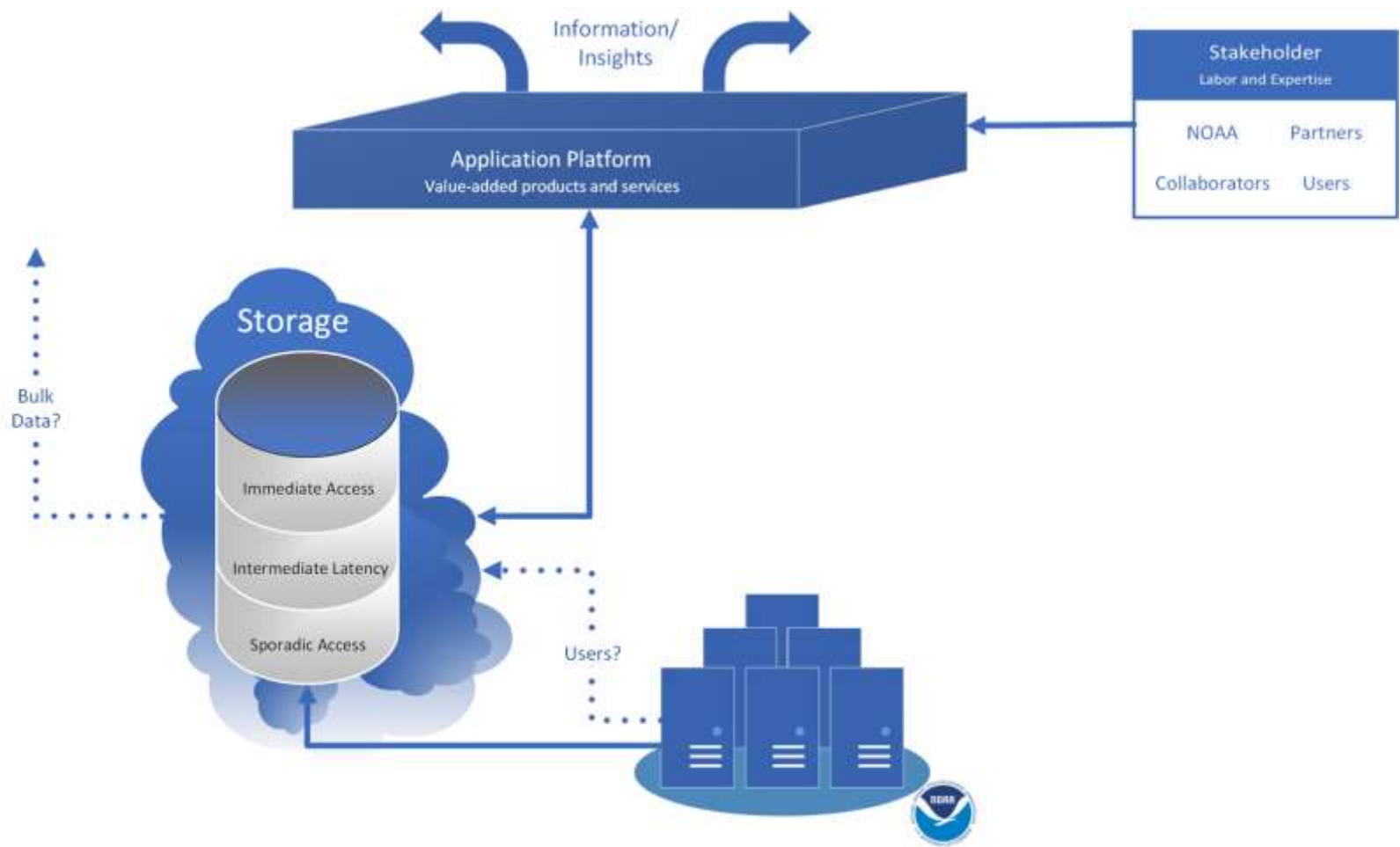
**Leverage the value of NOAA's data to increase their utilization**

# NOAA Datasets in Play...

- GOES 13 & 15
- AVHRR
- Disaster Response Platform\*
- Ocean Energy Platform\*
- Meteorological Assimilation Data Ingest System\* (MADIS)
- National Digital Forecast Database\*
- NOAA Port/SBN\*
- Anonymized Trawl Data
- Anonymized Observer Data
- Sea-level Rise\*
- Global Forecast System (GFS)
- Climate Data Records
- North American Multi-Model Ensemble\*
- Intermediate Modeling Products\*
- Pathfinder Sea Surface Temperature\*
- Ocean Bathymetry\*
- Multibeam Backscatter\*
- GOES-16
- NEXRAD L2
- NMFS Protected Species\*
- Essential Fish Habitat
- Rapid Refresh Modeling
- Global Historical Climatology Network - Hourly
- Global Historical Climatology Network - Daily
- Global Surface Summary of the Day
- International Comprehensive Ocean-Atmosphere Dataset (ICOADS)
- Filtered Alert Hub\*
- National Water Model
- Climate Forecast System - Version 2 (CFS v2)
- Fisheries Genomics/Meta-Genomics\*
- NMFS Commercial Landing Data\*
- VIIRS Night Lights Products\*
- Multi-Radar/Multi-Sensor (MRMS)\*
- Passive acoustic soundings\*

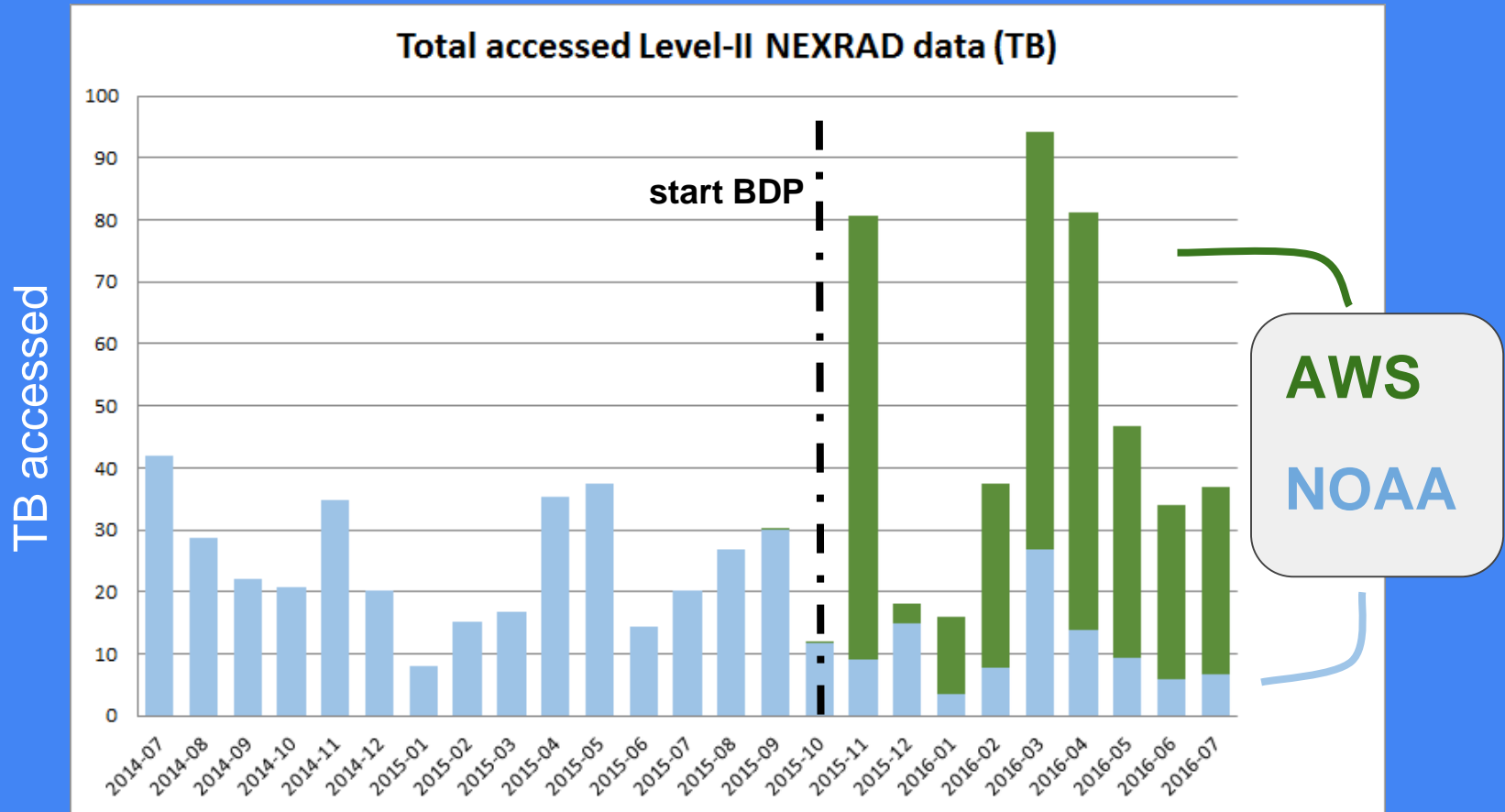
\* - Indicates activity underway

# Increased Usage of NOAA Data via BDP





# NEXRAD Weather Radar Data



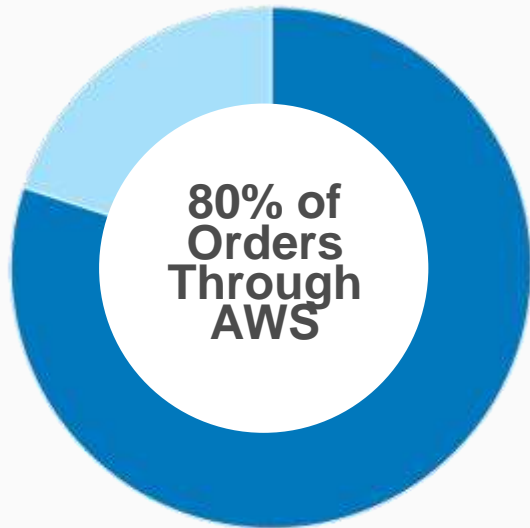
AWS: Oct '15 <https://s3.amazonaws.com/noaa-nexrad-level2> (1991+)

OCC: Jun '16 <http://occ-data.org/NOAANEXRAD/> (2015+) (S. Ansari et al, 2016)

# Example BDP Success Story

NEXRAD Level 2 Radar Data on AWS

**NOAA Wins**



■ AWS

■ NCEI

**AWS?**



**End User Wins**




■ AWS Job Time ~days  
■ Through NCEI ~Years

# Google NEXRAD Access

<https://cloud.google.com/blog/big-data/2017/06/visualization-and-large-scale-processing-of-historical-weather-radar-nexrad-level-ii-data>

As of June 15, 2017

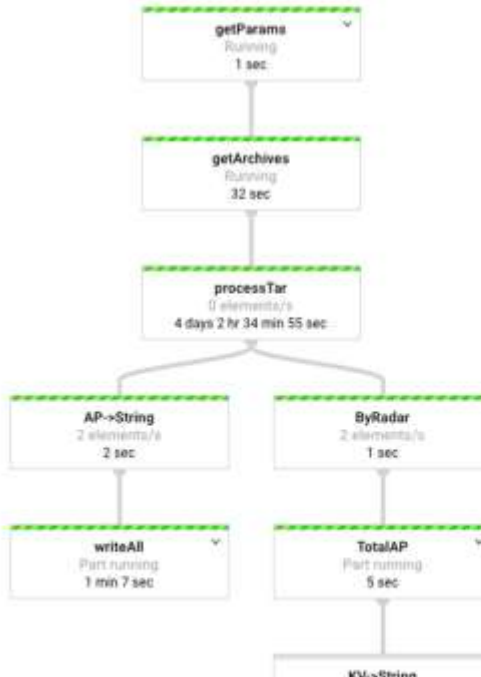
 Google Cloud Platform

[Why Google](#) [Products](#) [Solutions](#) [Launcher](#) [Pricing](#) [Customers](#) [Documentation](#) [Support](#) [Partners](#)

## Sample program to do large-scale analysis

While you can work with individual volume scans as shown above, one key benefit of having all the NEXRAD data immediately available on a public cloud is the ability to analyze long time periods of data at scale. Thanks to GCP's "serverless" approach to infrastructure, it's possible to do data processing, data analysis and machine learning without having to manage low-level resources.

[Cloud Dataflow](#), GCP's fully-managed service for stream and batch processing, allows you to write a data processing and analysis pipeline that will be executed in a distributed manner. The pipeline will autoscale different steps to run on multiple machines in a fault-tolerant way.




```
graph TD; getParams[getParams  
Running  
1 sec] --> getArchives[getArchives  
Running  
32 sec]; getArchives --> processTar[processTar  
0 elements/s  
4 days 2 hr 34 min 55 sec]; processTar --> APString[AP->String  
2 elements/s  
2 sec]; processTar --> ByRadar[ByRadar  
2 elements/s  
1 sec]; APString --> writeAll[writeAll  
Part running  
1 min 7 sec]; ByRadar --> TotalAP[TotalAP  
Part running  
5 sec]; writeAll --> KVString[KV->String  
Not started]; TotalAP --> KVString
```

### Job summary

Job name	appipeline-vlakshmanan-0522205052-597f064b
Job ID	2017-05-22_13_50_54-9553301786727587053
Job status	Running
<button>Stop job</button>	
SDK version	Google Cloud Java 2.0.0
Job type	Batch
Start time	May 22, 20
Elapsed time	53 min 26 s

### Autoscaling

Workers	15
Current state	Worker pool



May 22, 2:00 PM

Current workers: 15 Target workers

[See more history](#)

Google Cloud Platform Select a project

Storage Browser

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER

Filter by prefix

Buckets / gcp-public-data-nexrad-0 / 2015 / 04 / 01 / KABR

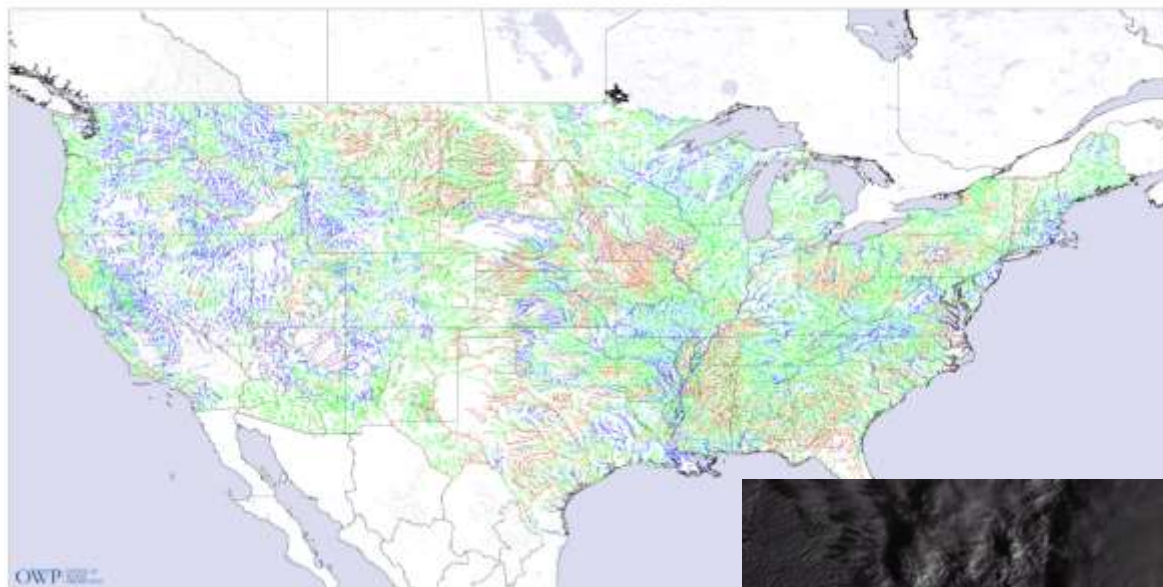
Name	Size
NWS_NEXRAD_NXL2DP_KABR_20150401000000_20150401005959.tar	17.6 MB
NWS_NEXRAD_NXL2DP_KABR_20150401010000_20150401015959.tar	25.45 MB
NWS_NEXRAD_NXL2DP_KABR_20150401020000_20150401025959.tar	26.66 MB
NWS_NEXRAD_NXL2DP_KABR_20150401030000_20150401035959.tar	28.64 MB
NWS_NEXRAD_NXL2DP_KABR_20150401040000_20150401045959.tar	20.91 MB
NWS_NEXRAD_NXL2DP_KABR_20150401050000_20150401055959.tar	30.26 MB
NWS_NEXRAD_NXL2DP_KABR_20150401060000_20150401065959.tar	29.36 MB
NWS_NEXRAD_NXL2DP_KABR_20150401070000_20150401075959.tar	31.68 MB
NWS_NEXRAD_NXL2DP_KABR_20150401080000_20150401085959.tar	22.21 MB
NWS_NEXRAD_NXL2DP_KABR_20150401090000_20150401095959.tar	21.12 MB
NWS_NEXRAD_NXL2DP_KABR_20150401100000_20150401105959.tar	20.54 MB

# Ongoing and Upcoming Efforts

## National Water Model Streamflow Anomaly Guidance

Analysis valid for 2017-06-01 19:00:00 UTC

Model initialized at 2017-06-01 18:00:00 UTC



National Water Center: <http://water.noaa.gov/tools/nwm-image-viewer>

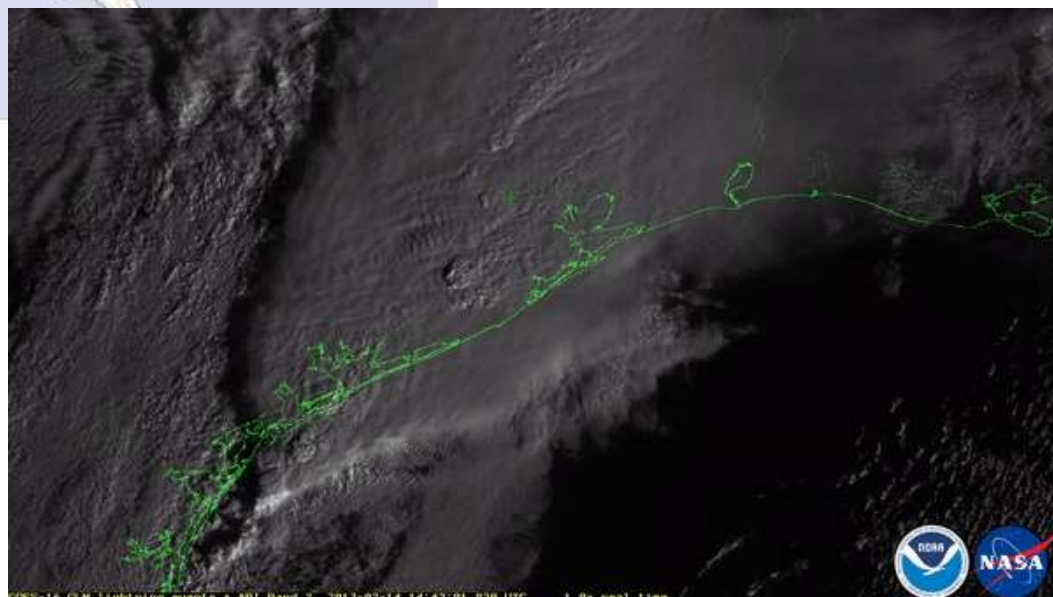
## National Water Model:

- 23-year reanalysis
- Real-time forecast

## GOES-16:

- Now: L1b ABI Products
- Began July 12, 2017
- Provisional status
- Soon: L2 products (GLM)

NOAA NESDIS: <https://www.nesdis.noaa.gov/content/flashy-first-images-arrive-noaa%E2%80%99s-goes-16-lightning-mapper>



# NCEI User Requests - By Sector





# NCEI User Requests - By Theme

Precipitation 37%	Wind 11%	Snow 8%
Temperature 30%	Storm 8%	Pressure 3%
		Humidity 2%
		24

# Why is NOAA interested in this?

## NCEI User Profiles

% of Users	Typical User	Requested Data Type	Preferred Format	How Much?	How Often?	System Impact
70	General business, media, public	Qualitative	Point+Click, visualization assessment	Low	High	Low
15	Researchers, business consultants	Quantitative	Digital downloads	High	Low	High
15	Value-added Providers (database scrapers)	Quantitative	Machine to machine downloads	Low	High	High

# Big Data Project Methodology

01

## Business Discovery

CRADA Collaborators & any Third-Party Partners work together to identify datasets of interest & develop business cases

02

## Initial Technical Discussion

Develop a strategy for data delivery from NOAA to BDP Collaborators

03

## In-Depth Data Discussions

Engage NOAA SMEs, BDP Collaborators for technical interchanges

04

## Product Development

Collaborators and their Partners create services

- ◆ Develop markets & financial opportunities based on NOAA data
- ◆ Generate revenue and profits

05

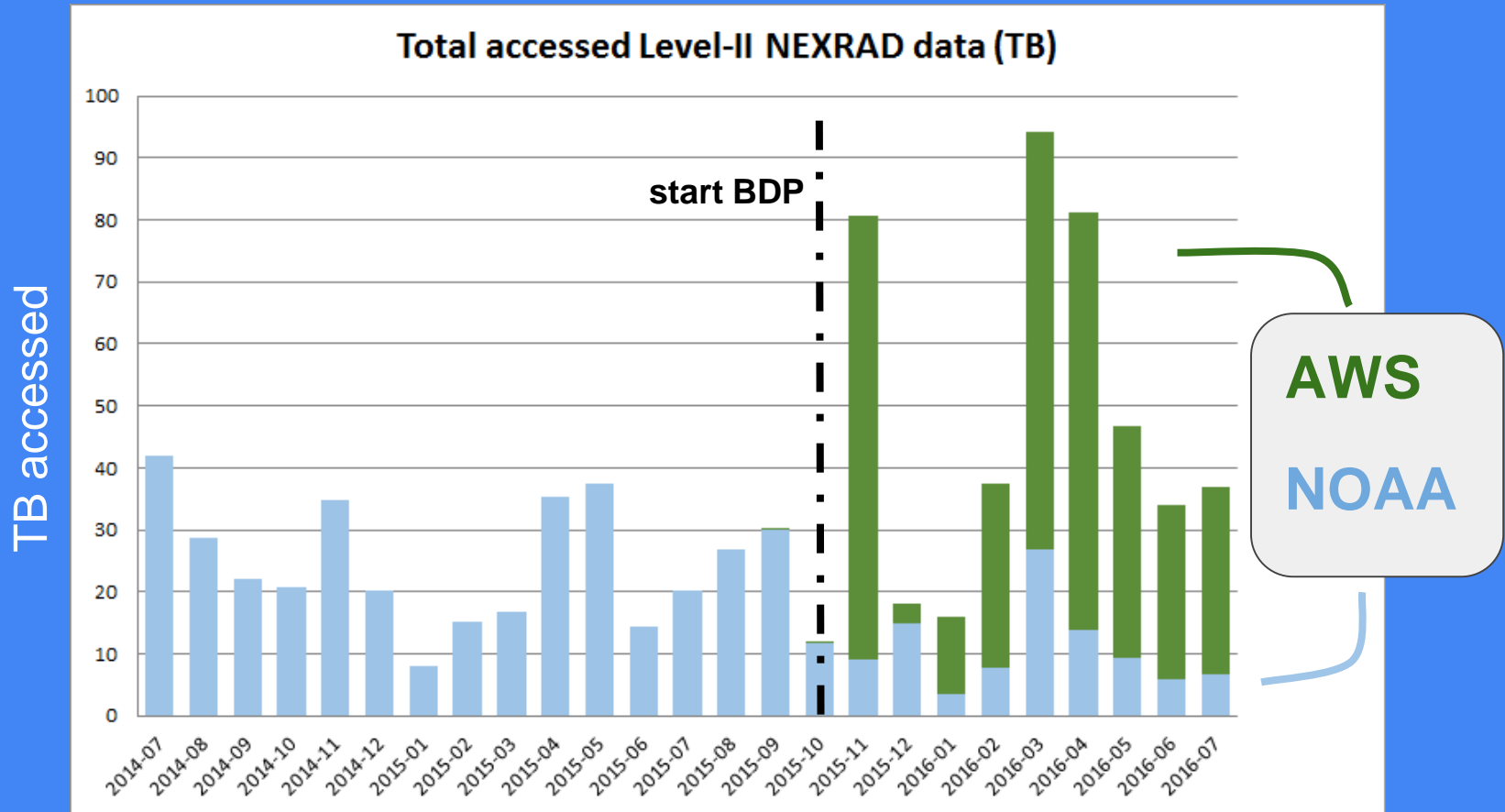
## Augmented NOAA Services

NOAA continues all of it's existing data services

- No interruption of existing services to customers, but new options
- BDP activities are an augmentation of existing services



# NEXRAD Weather Radar Data



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)