Update on NOAA's Big Data Project

Dr. Edward Kearns
NOAA Chief Data Officer
National Oceanic and Atmospheric Administration

August 24th, 2017

Acknowledgements

Many thanks to:

- BDP Core Team: Andy Bailey, Shane Glass, Jeff de la Beaujardiere, Tony LaVoi, Jay Morris, Derek Parks
- NOAA: Brian Eiler, Zach Goldstein, Dave Michaud, Glenn Tallia, , Derek Hanson, Kate Abbott, Amy Gaskins*, Alan Steremberg*, Maia Hansen*, Steve Ansari, Steve Del Greco*, Brian Nelson, Carlos Rivero*, Ken Casey, Rich Baldwin, Ed Clark, Brian Cosgrove, Donna McNamara, Chris Sisko, Nathan Wilson, Mark Brady*
- NC State University / CICS-NC: Otis Brown, Scott Wilkins, Jonathon Brannock, Lou Vazquez, Scott Stevens, Paula Hennon, Andrew Buddenberg, Angel Li

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

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

Introduction to NOAA's new CDO

- NOAA's new Chief Data Officer position was created in 2017 as an SES position within the Office of the Chief Information Officer

- Duties include creating/evolving an overall NOAA Data Strategy, developing NOAA Data Policy, increasing data usability, and data issues with other US Agencies and Industry.
  - work across all of NOAA's Line Offices and Staff Offices
  - "mission" data and business data

- Who am I? Ed Kearns, have been with NOAA for 9 years at NESDIS/NCEI.
  - Archive, climate data records, satellite calibration, data systems
  - Univ of Miami/RSMAS, National Park Service/Everglades, NOAA/NDBC
  - Ph.D. Physical Oceanography (1996, URI/GSO)
    https://www.linkedin.com/in/edwardkearns/

Why is NOAA Interested in Open Data Partnerships?

- NOAA data are **increasingly popular and valuable**
- NOAA struggles to keep up with public demand
  - **Budgets** for capacity and security: **Static**
  - **Costs** for data access demand: **Increasing**
- NOAA wants to learn about collaborative solutions
  - Promote use, improve data access
  - Enable new economic opportunities

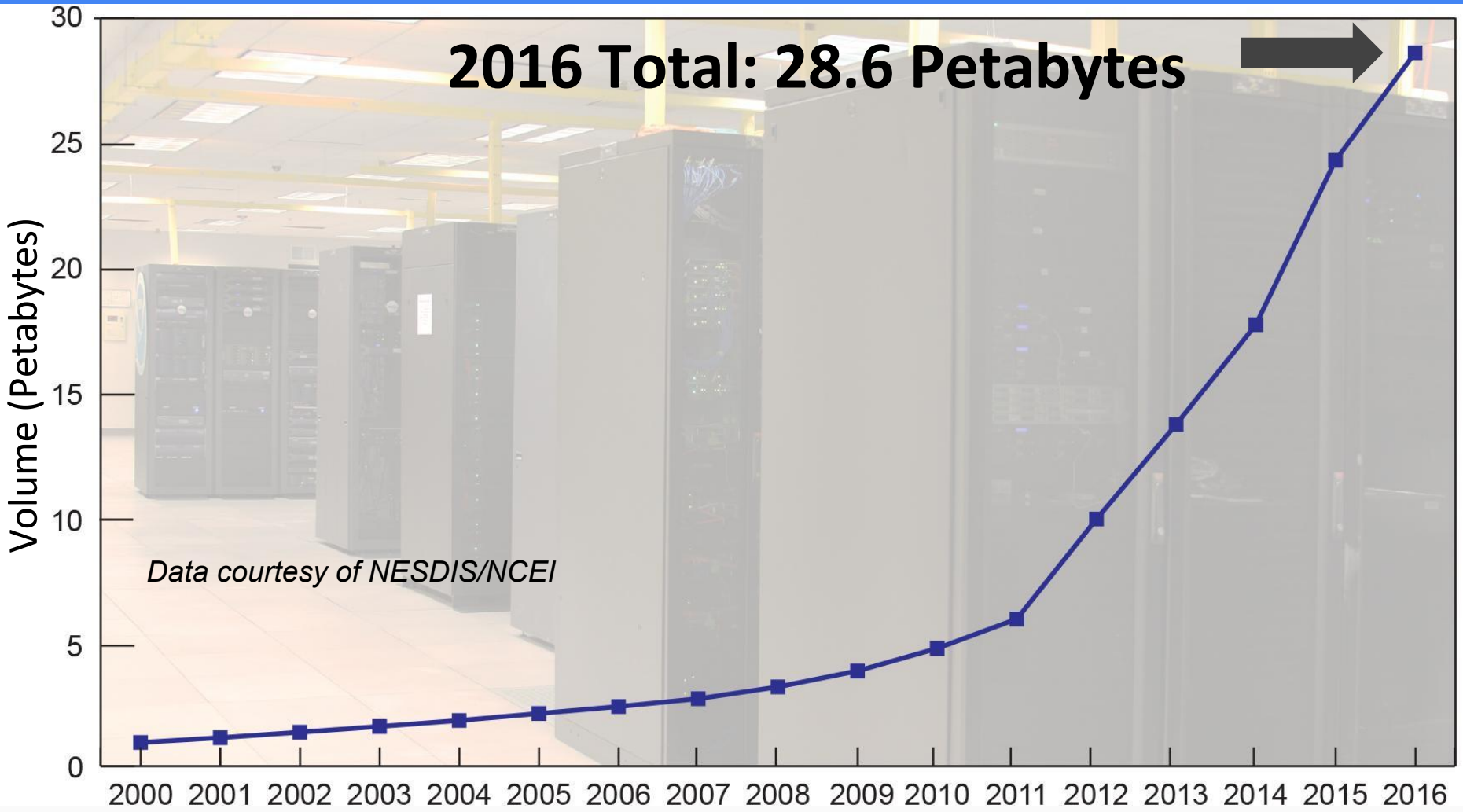
Why is NOAA interested in this?

NCEI User Profiles

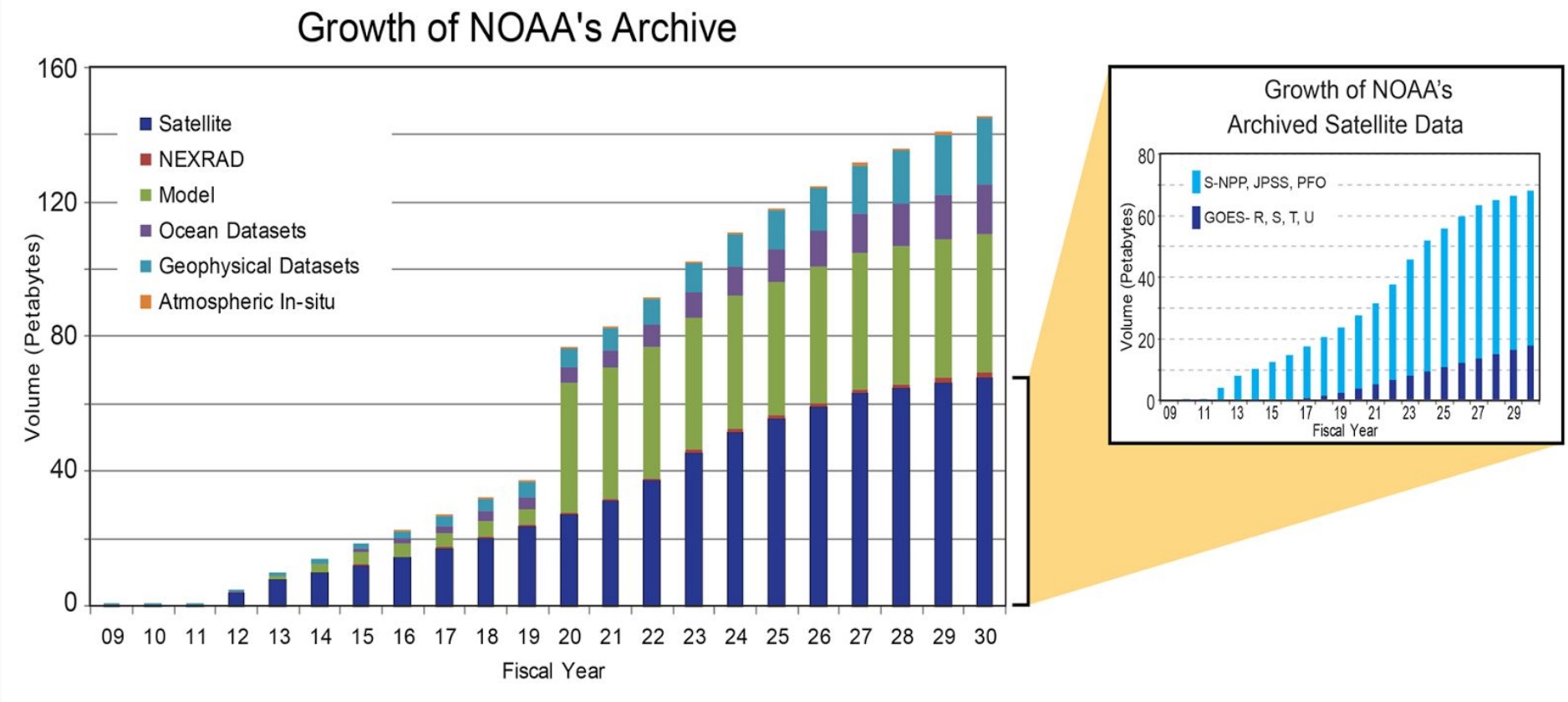
% of Users	Typical User	Requested Data Type	Preferred Format	How Much?	How Often?	System Impact
70	General business, media, public	Qualitative	Point+Click, visualization assessment	Low	High	**Low**
15	Researchers, business consultants	Quantitative	Digital downloads	High	Low	**High**
15	Value-added Providers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

NOAA/NCEI's Environmental Data Archive

Increasing volumes from station, model, radar, & satellite data

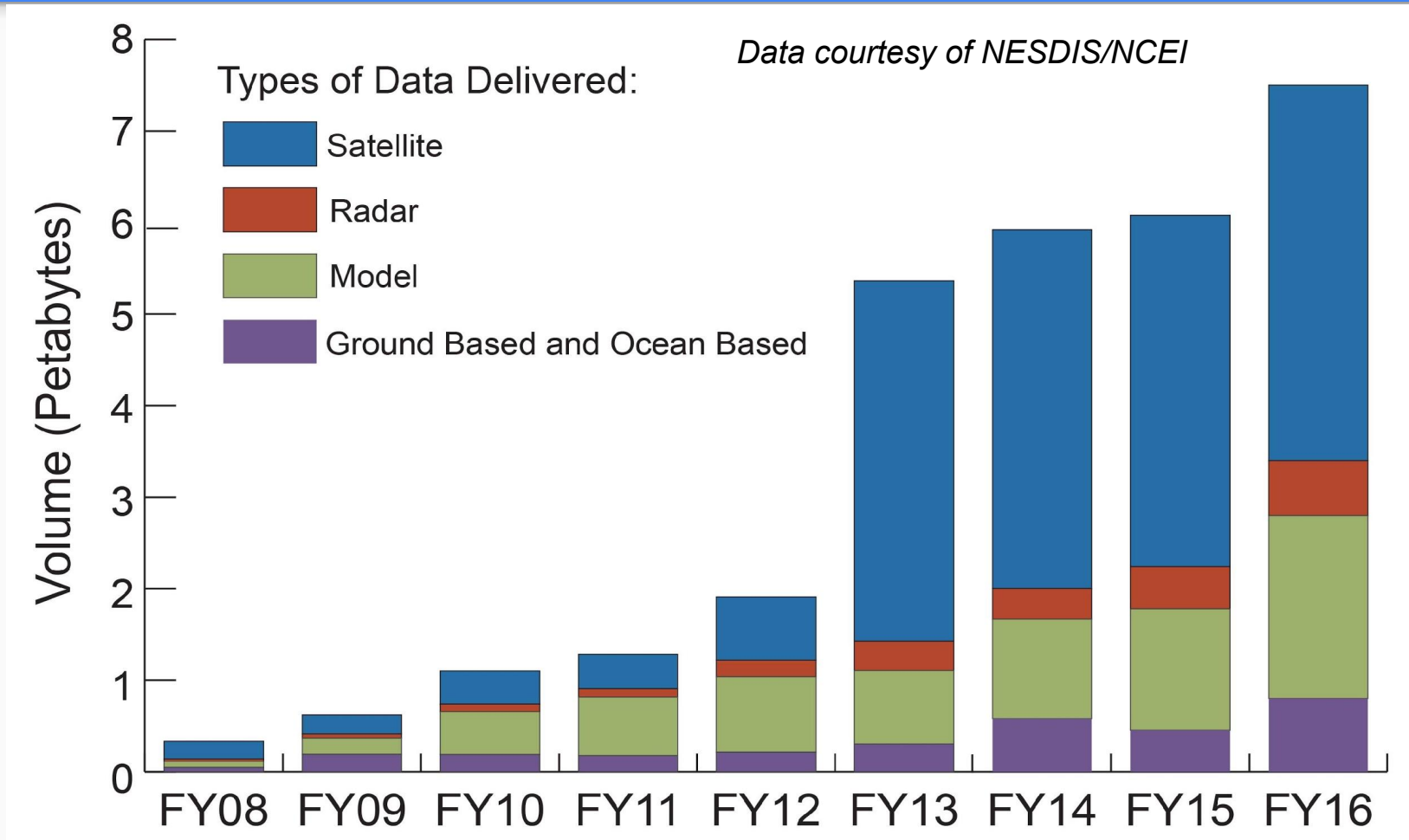


Data Volumes Rapidly Increasing

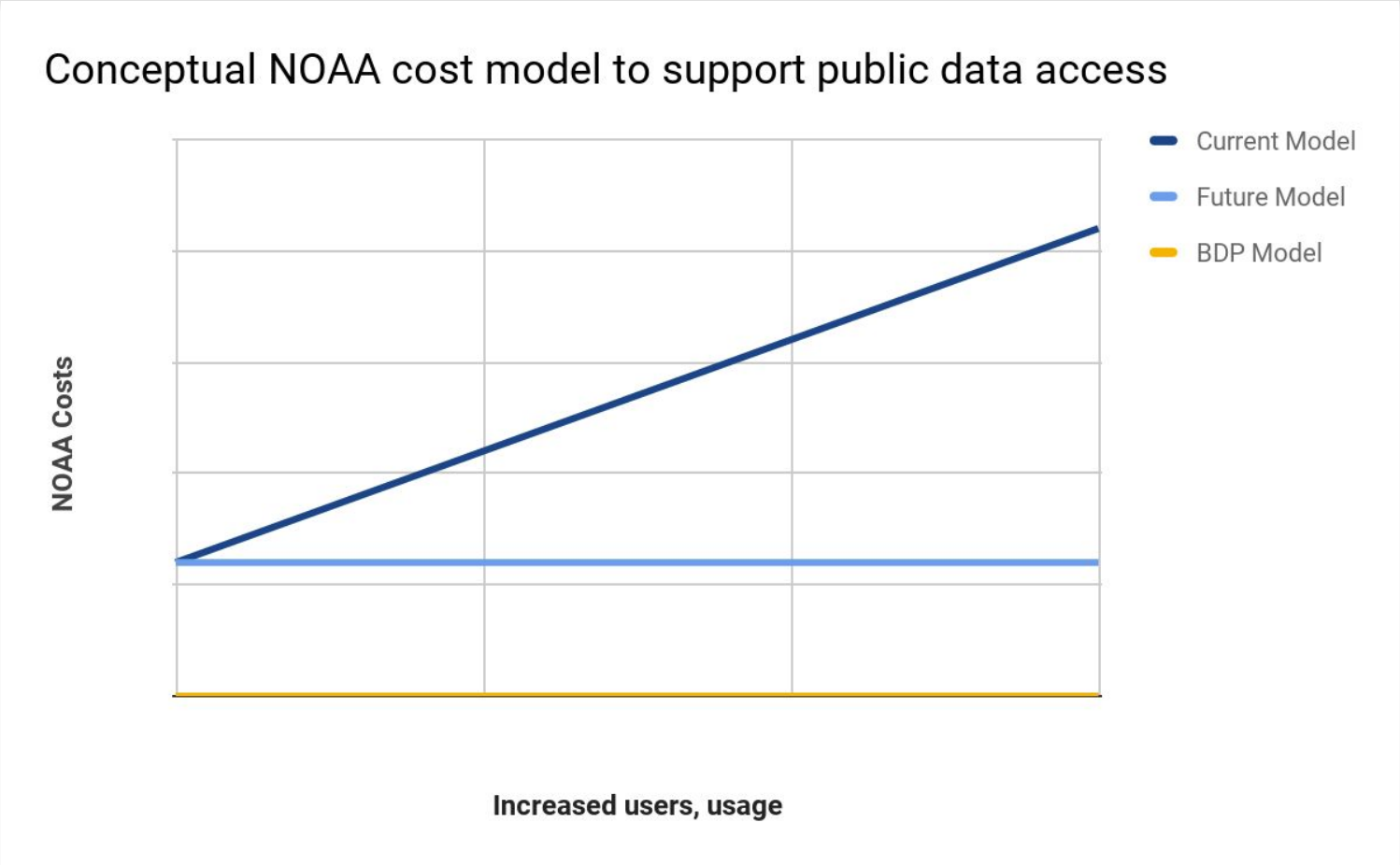


Data courtesy of NESDIS/NCEI

Demand for Access Increasing



Need to investigate how to change NOAA's financial model for public open data access



The Big Data Project

- Cooperative Research and Development Agreements
  - 5 separate but identical 3-year agreements

- The Basics of the Agreement
  - Industry provides access to NOAA's data to all
    - Original data are provided freely
    - Monetize services and Products based on data
  - NOAA provides data and expertise

- Combines 3 powerful resources based on NOAA's open data:
  - NOAA's subject matter expertise
  - Industry's data storage and access expertise
  - Cloud's scalable and on-demand processing capability

BDP Data Access Strategy

Leveraging Industrial Partners' Capabilities

Augment

Add Capabilities

Amplify

Add Capacity

Bring Processing to the Data

Big Data Project:

Better, Cheaper, More Secure Access

- Wider access to NOAA's open data
  - More secure (copies of data accessed from non-federal systems)
- Cost Avoidance for public data access
  - Most popular datasets bring largest burden on NOAA systems
    - Are also the best candidates for this partnership model
- Better Level of Service to customers
  - Users can utilize data faster without downloading it
- **This is not *just* about open data access**
  - **Can accelerate data utilization…**
  - **...and thus improve societal impacts and business opportunities**

All existing NOAA service outlets remain. Big Data Project (BDP) offers alternatives and advantages to explore

Augmentation, not replacement

Fair and Level Access

NOAA will offer equal access to the data for all collaborators

CRADA Collaborators Responded to RFI

No Net Cost to Taxpayers

Collaborative Research And Development Agreement (CRADA)

As part of the CRADA, NOAA may recover costs for new or supplemental efforts

Data remains free and open

Monetize Value-added products and services

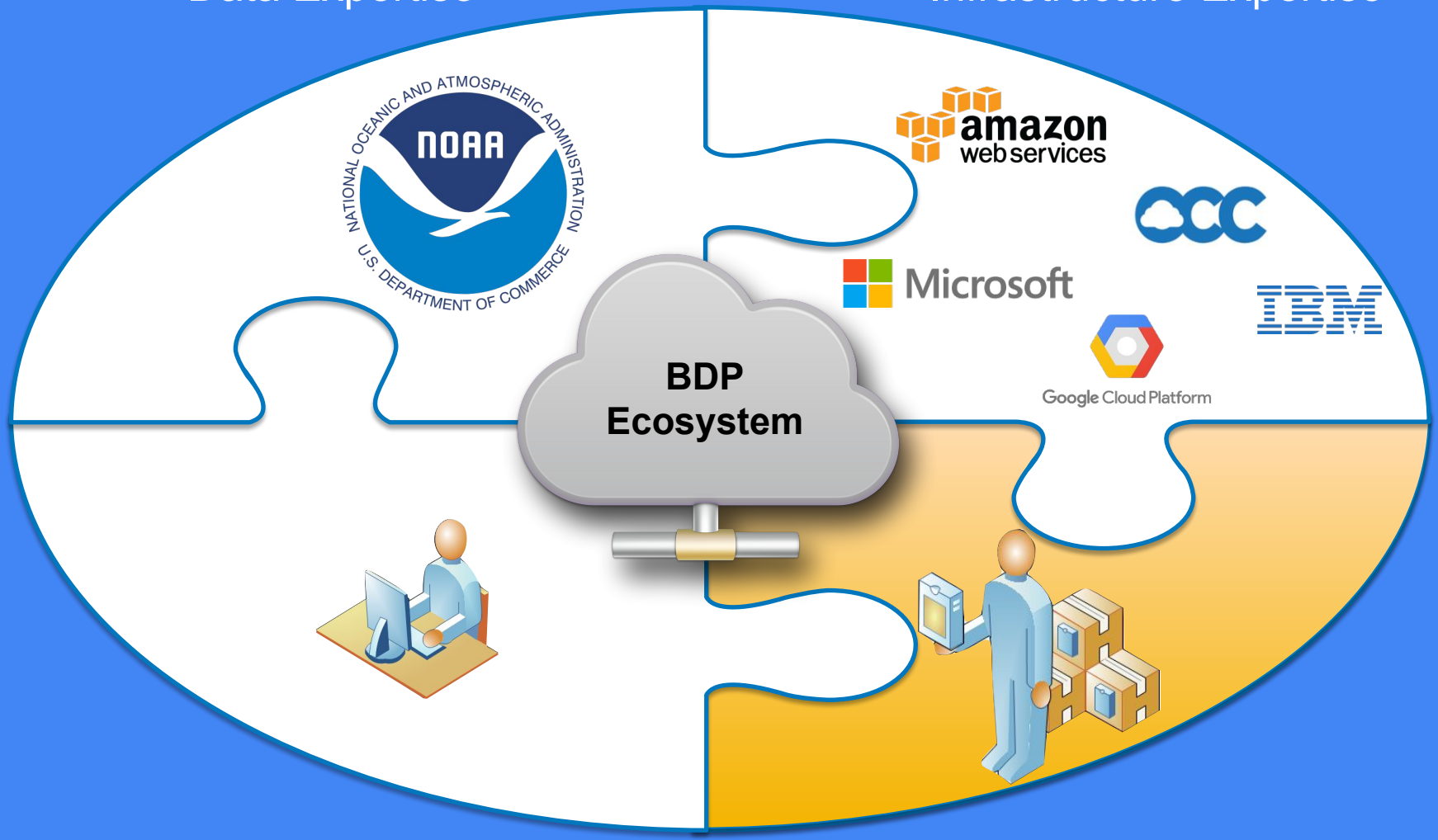
Original NOAA data can be downloaded for free through collaborators. Collaborators may recover costs associated with data acquisition

Collaborators generate revenue when 3rd parties utilize the data. Collaborators may charge for value-added services and products

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

NOAA
Data Expertise

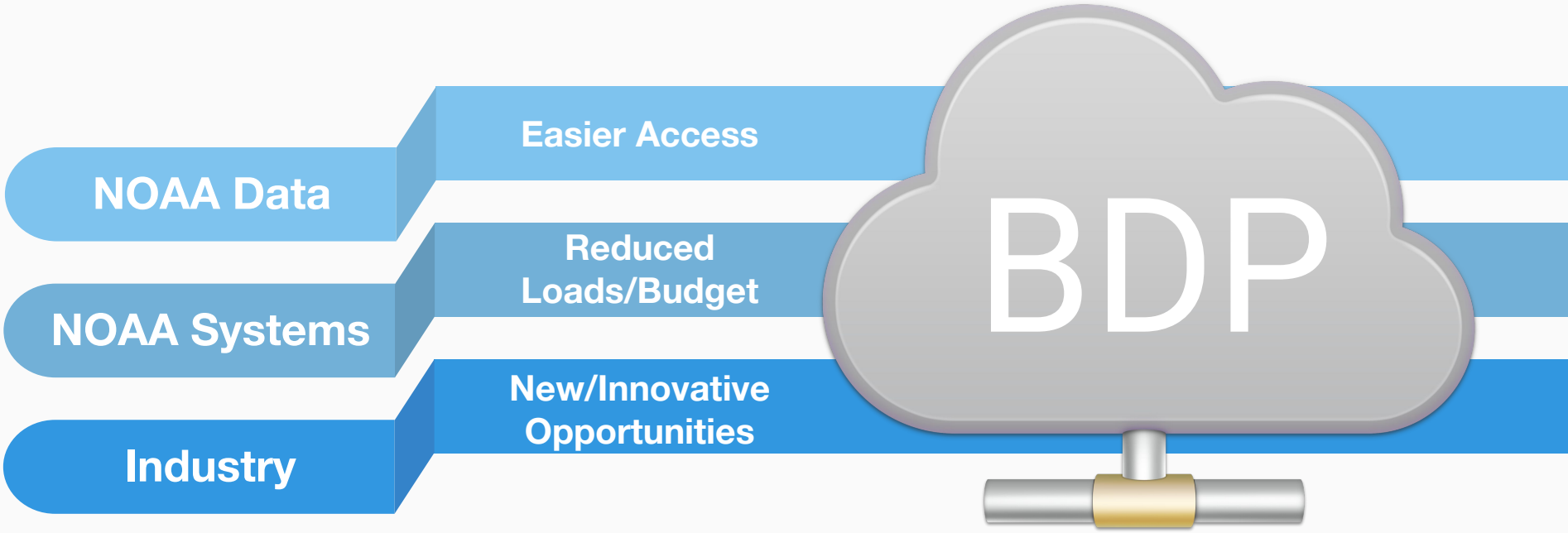
CRADA Collaborators
Infrastructure Expertise



End User
Wider Consumer Community

Third Party Partner
Value-Added Services

Tangible BDP Benefits



Big Data Project Collaborators' Data Offerings

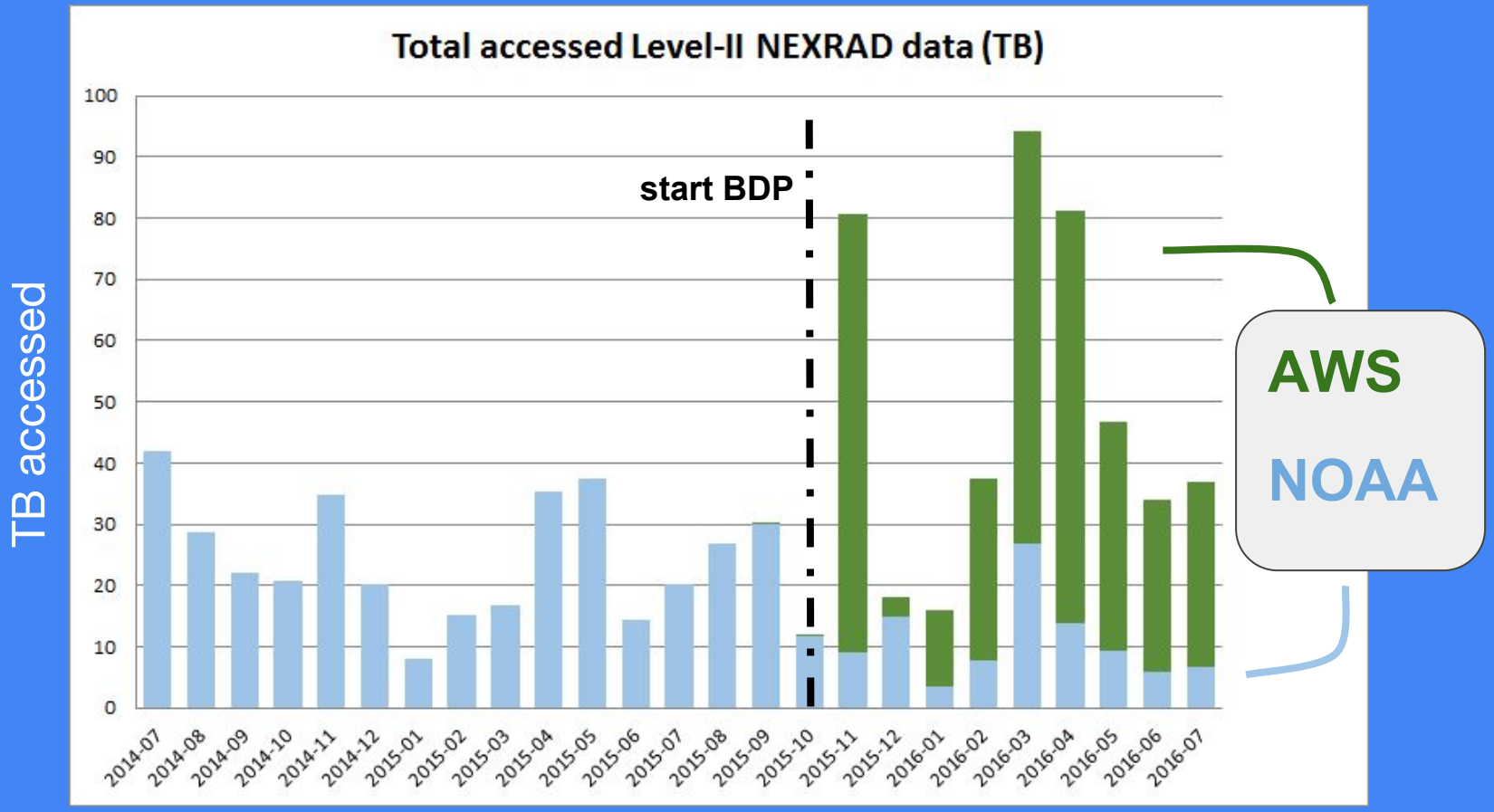
- **AWS**
  - https://aws.amazon.com/noaa-big-data/
- **Google Cloud Platform**
  - https://cloud.google.com/bigquery/public-data/ *(see NOAA listings on left)*
- **IBM**
  - https://noaa-crada.mybluemix.net/
- **Microsoft**
  - No public services released yet
- **Open Commons Consortium**
  - http://edc.occ-data.org/

AWS and NEXRAD

BDP Success Story

- **2nd Highest US National Observation Value**
  - National Plan for Civil Earth Observations (2014)
- **Entire** 88D Archive transferred to AWS and OCC 2015
  (as well as two others who haven't made their services public)
- **Options:** NOAA Redirects to BDP Collaborators' services
- Single access point for **archived** and **real-time data**
- 3rd parties **- Climate Corp and Unidata-** were key to success

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, 2017)*

Example BDP Success Story

NEXRAD Level 2 Radar Data on AWS

NOAA Wins

AWS?

End User Wins

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

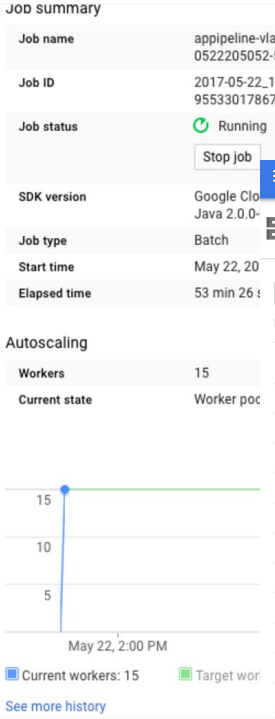
Google Cloud Platform

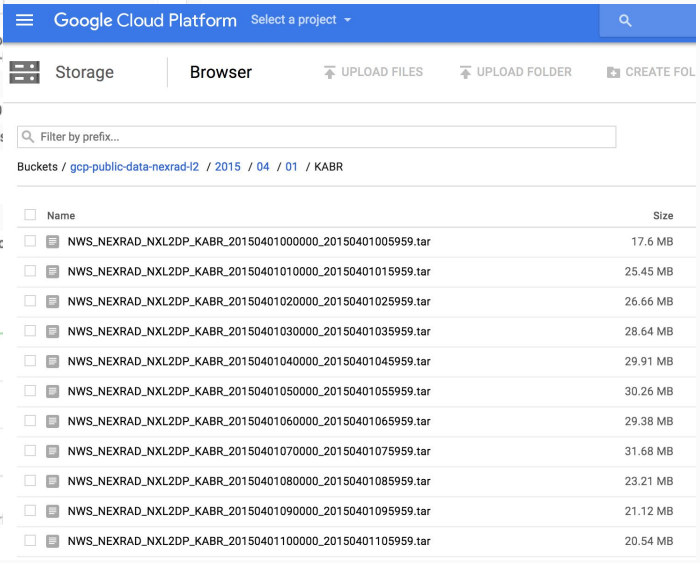
Why Google | Products | Solutions | Launcher | Pricing | Customers | Documentation | Support | Partners

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.





As of June 15, 2017

GCP and CDI/PReP

A BDP Success Story

Climate Data Online

Climate Data Online (CDO) provides free access to NCDC's archive of global historical weather and climate data in addition to station history information. These data include quality controlled daily, monthly, seasonal, and yearly measurements of temperature, precipitation, wind, and degree days as well as radar data and 30-year Climate Normals. Customers can also order most of these data as certified hard copies for legal use.

Browse Datasets
Browse documentation, samples, and links

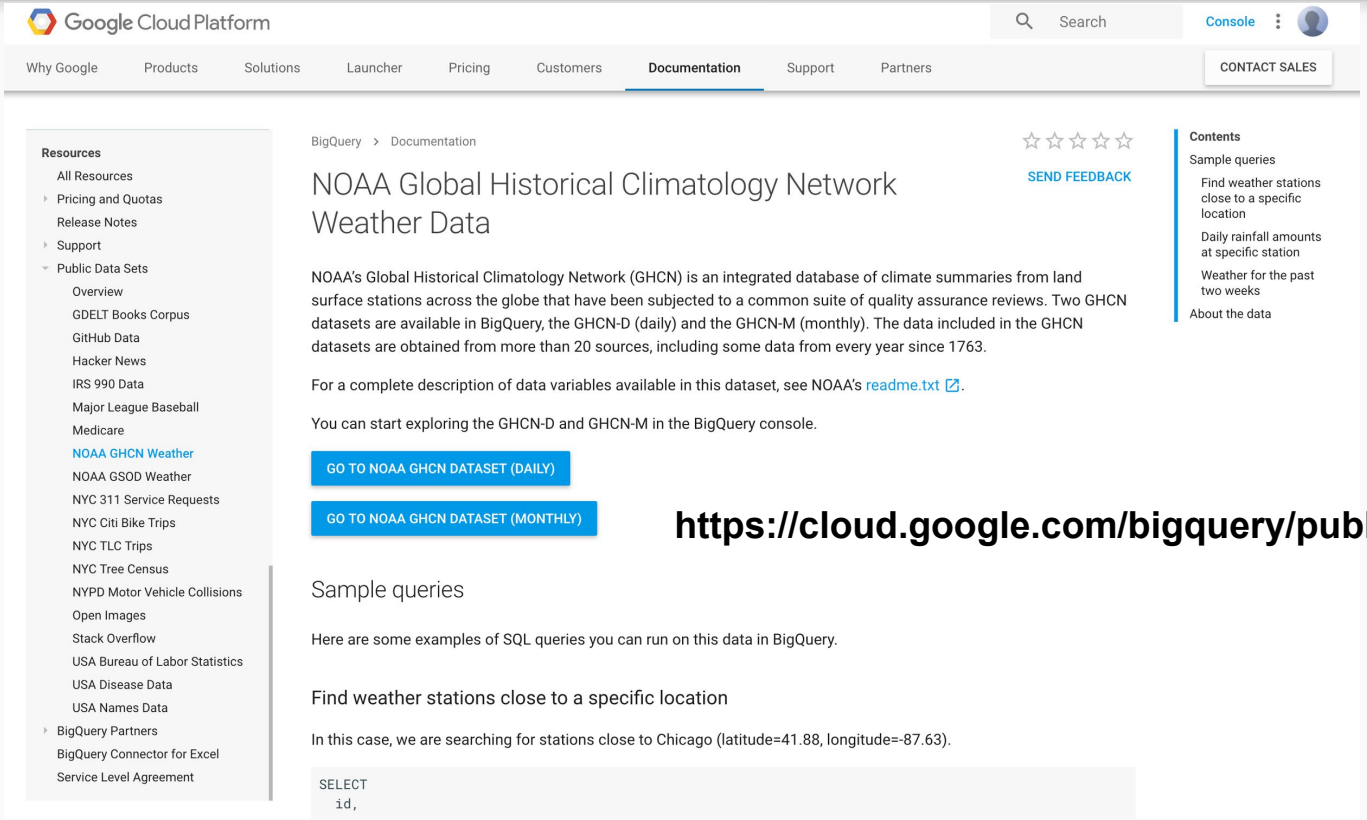
Certify Orders
Get orders certified for legal use (requires payment)

Check Status
Check the status of an order that has been placed

Find Help
Find answers to questions about data and ordering

NOAA NCEI: https://www.ncdc.noaa.gov/cdo-web/

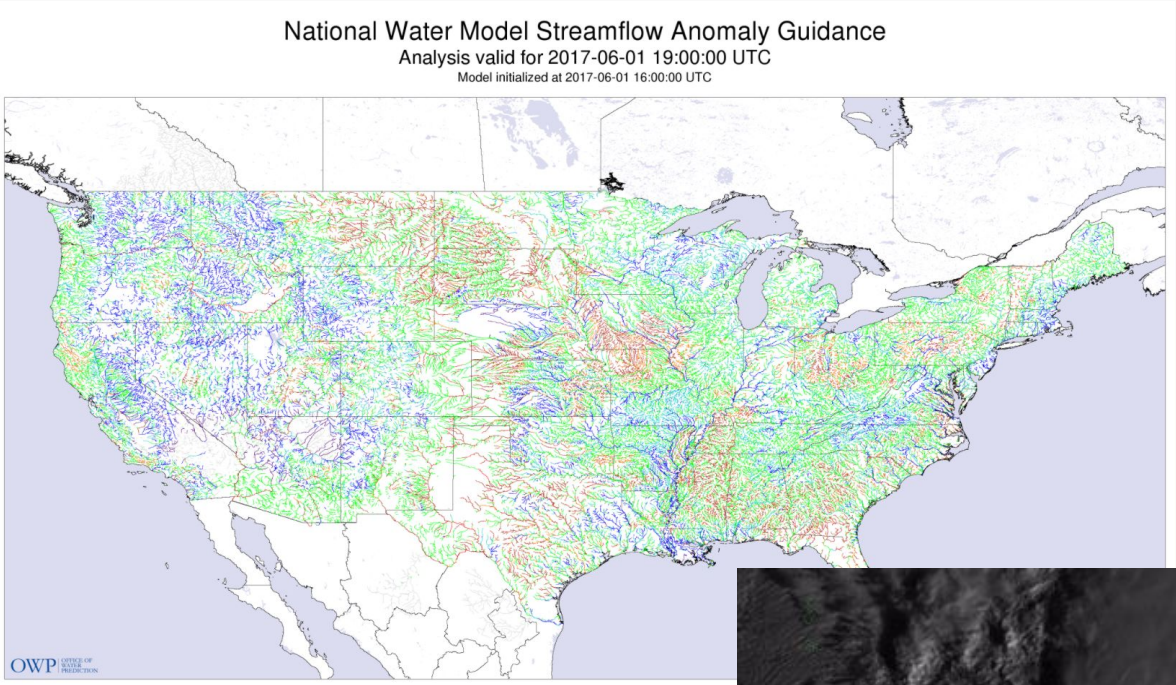
Google Cloud Platform Example



https://cloud.google.com/bigquery/public-data/noaa-ghcn

- **1.2 PBs** of climate and weather data accessed through Google BigQuery, in **4 months**
  - Without "trying" - not advertised yet
  - Joins, joins, joins
  - 30-100x of NOAA deliveries in that time
- Images in Google Earth Engine
  - GOES-16 (June 2017)
  - National Water Model data
  - Weather and Climate model output
  - Climate data records

Ongoing and Upcoming Efforts



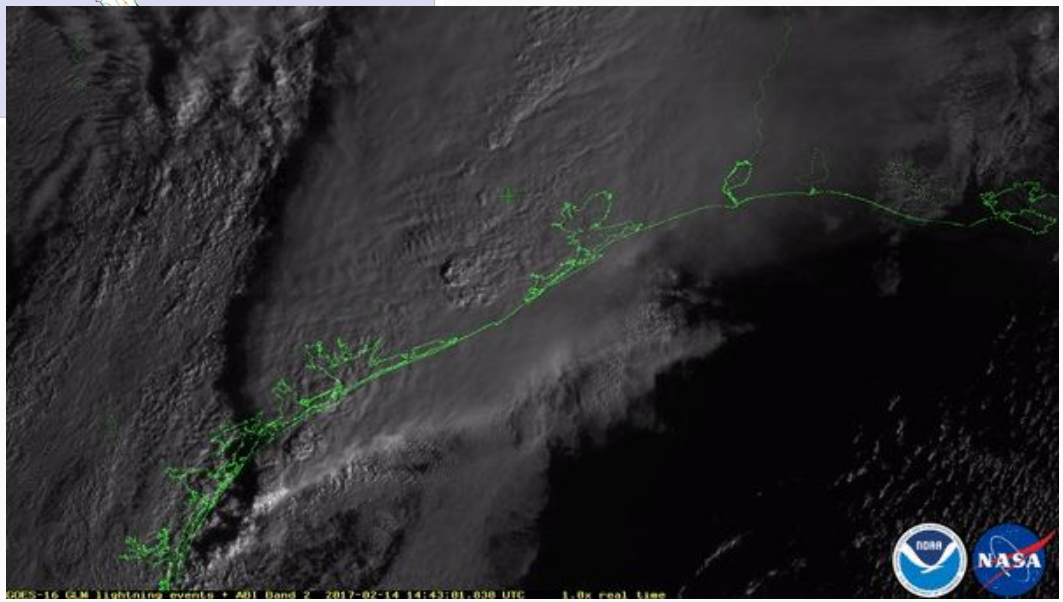
National Water Center:
http://water.noaa.gov/tools/nwm-image-viewer

National Water Model:

- 23-year reanalysis
- Real-time forecast

GOES-16:

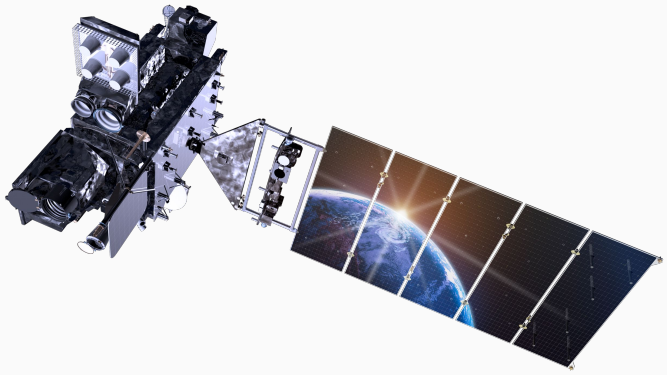
- Now: L1b ABI Products
- Began July 12, 2017
- Provisional status
- Soon: L2 products (GLM)

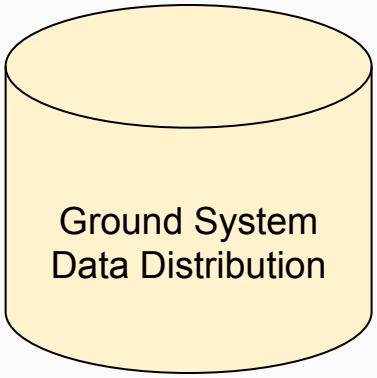


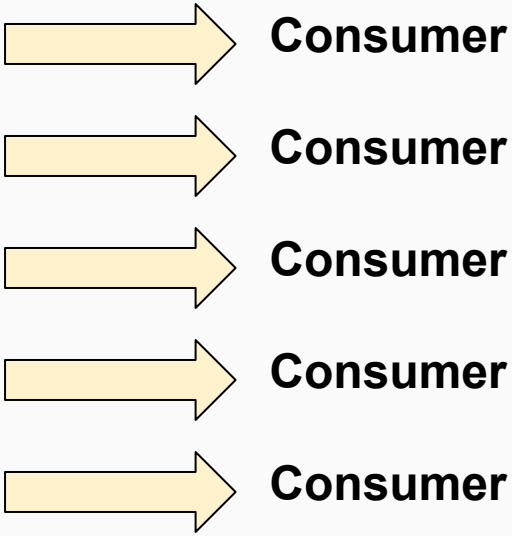
NOAA NESDIS:
https://www.nesdis.noaa.gov/content/flashy-first-images-arrive-noaa%E2%80%99s-goes-16-lightning-mapper

Traditional Satellite Data Internet Access Strategy

One-to-One Model

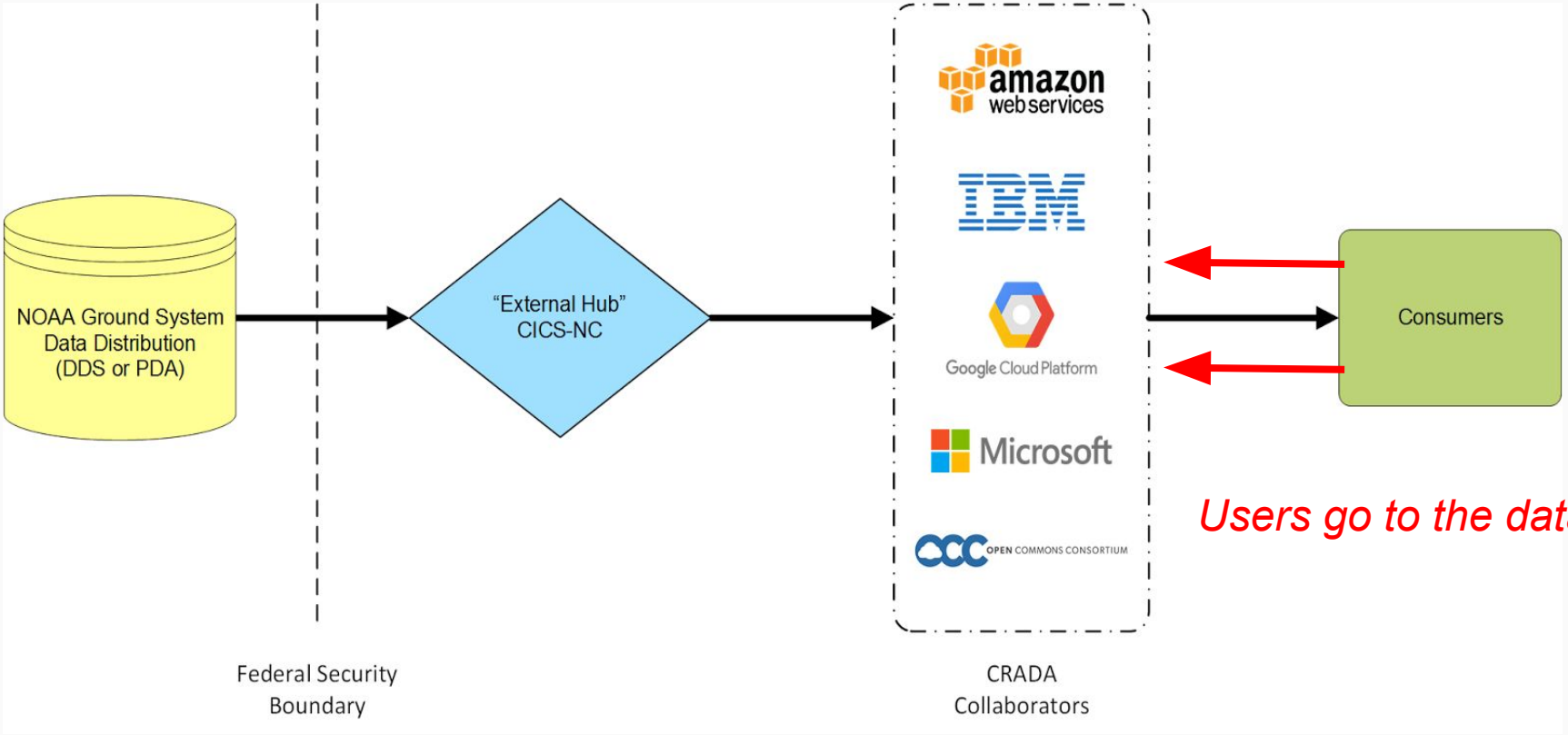






Big Data Project Satellite Data Access Demo Activity

One-to-Many Model for GOES-16 data



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**

NOAA is seeking Consumers' feedback

- Are the types of data access and services provided by the BDP and Collaborators meeting your needs?
- Does the BDP approach make things easier on the user?
- Encourage communications with the Collaborators
  - Help shape the services that you need
- Seek feedback from NOAA on the BDP

Big Data Project and Open Data Challenges

- How do we understand the Big Data market?
  - All NOAA's data commercially-viable in this model?
  - How can we more systematically select datasets?

- How to best transfer and steward numerous large, complex datasets?
  - How to ensure data authenticity?
  - Real-time, e.g. satellites, weather observations, coastal data
  - Retrospective, e.g. climate models and observations, fisheries

- What comes next, after the CRADA expires?
  - Spin off new agreements or partnerships as April 2018 nears?
  - Have we learned enough yet? Extend CRADAs to learn more?

Need engagement & feedback from NOAA, Collaborators, users

Summary

- NOAA is collaborating with industry through the Big Data Project CRADA to learn how to make NOAA's data **more easily and widely usable,** in a cost-effective manner.

- The BDP experience is showing that modern platforms provide:
  - **Higher Levels of Service** to the customer
  - **Reduced loads on NOAA** access systems that reduce cost

- A new public access paradigm for NOAA data, at **reduced cost**?
  - NOAA provides minimal services, **fair-and-level** access
  - Cloud collaborators add **scalable capacity & capabilities**
  - **NOAA stewards its data throughout** the cloud landscape

- Applications can be developed **faster**, using **less bandwidth**
  - Authoritative data are co-located with the processing capacity

Discussion

ed.kearns@noaa.gov

#NOAABigData

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

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

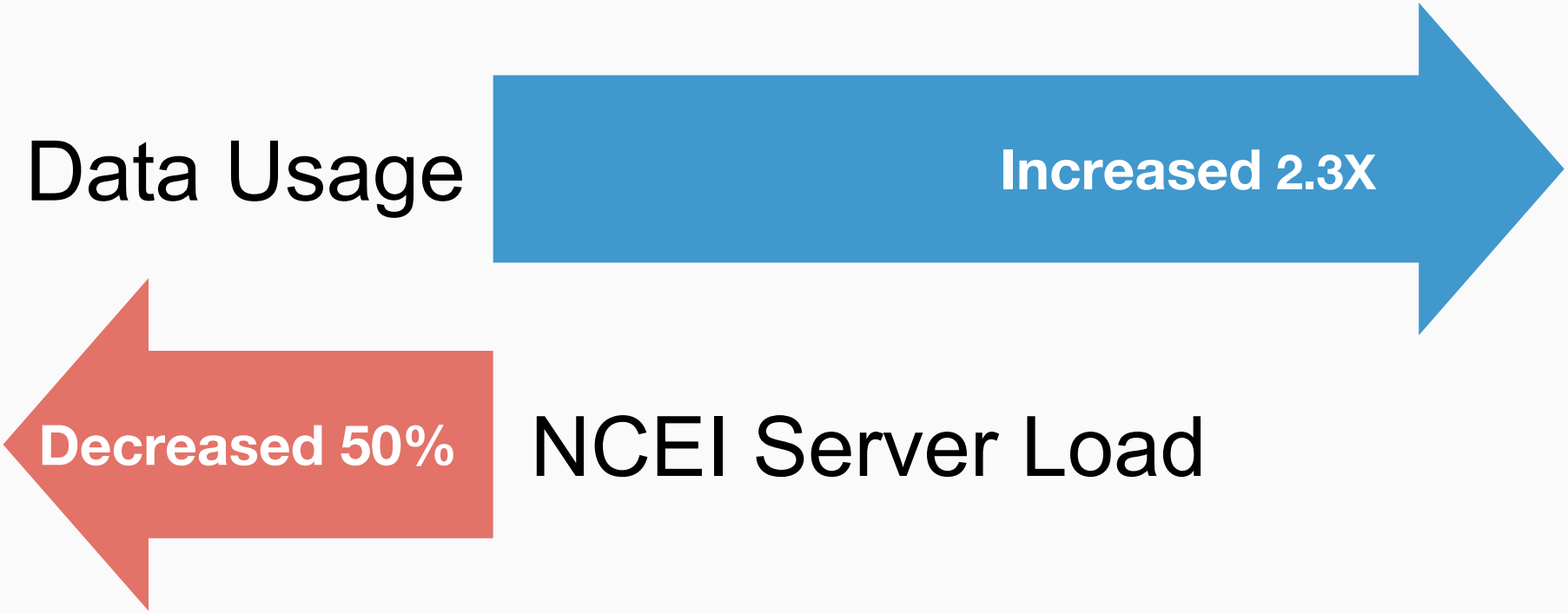
Augmented NOAA Services

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

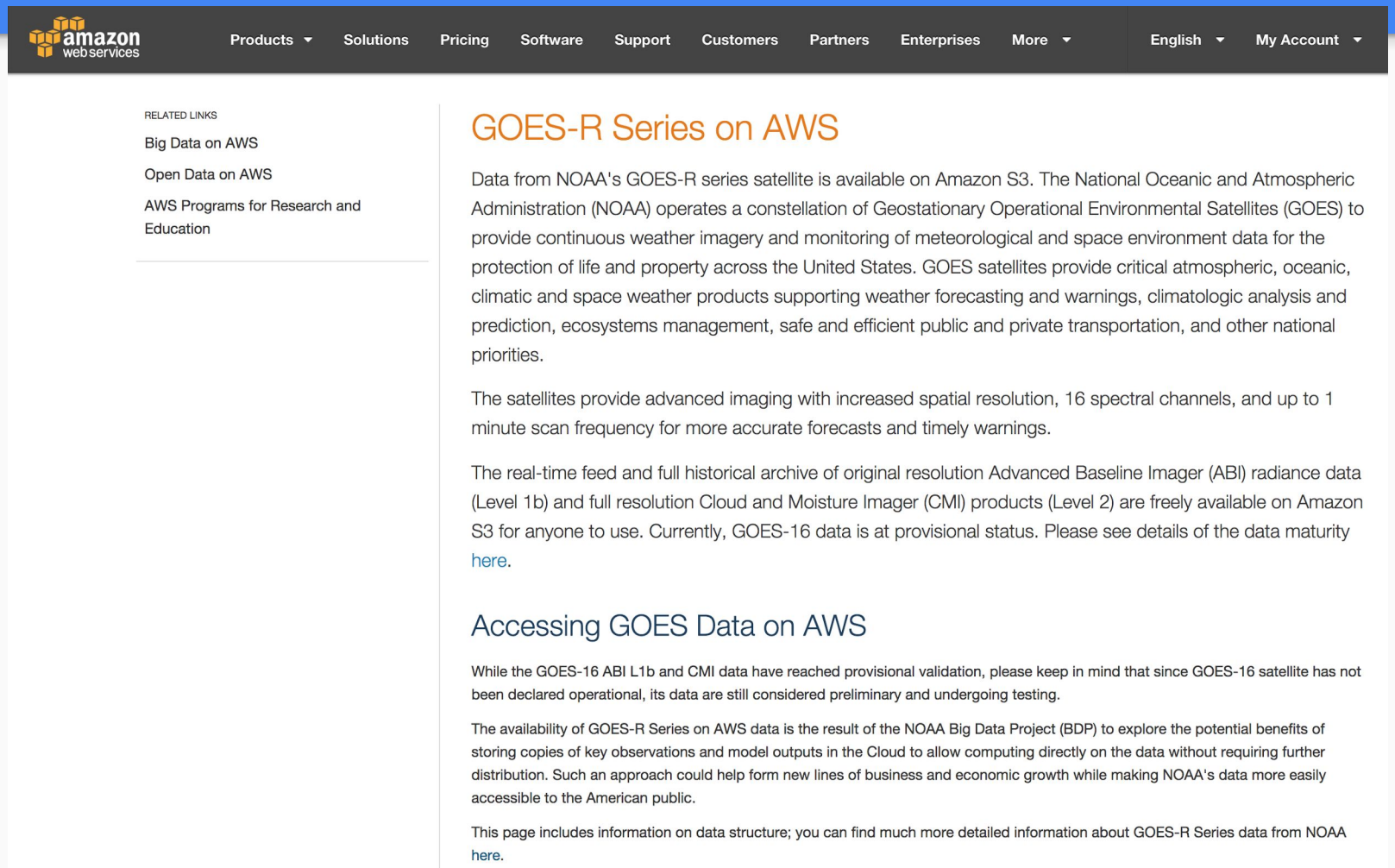
AWS and NEXRAD

A BDP Success Story



AWS GOES-16

https://aws.amazon.com/public-datasets/goes/



AWS GOES-16

https://aws.amazon.com/public-datasets/goes/


```
aws s3 ls noaa-goes16

aws s3 cp s3://noaa-goes16/<Product>/<Year>/<Day of Year>/<Hour>/<Filename>
```

Subscribing to GOES Data Notifications

We have set up public Amazon SNS topics that create a notification for every new object added to the Amazon S3 buckets for GOES on AWS. To start, you can subscribe to these notifications using Amazon SQS and AWS Lambda. This means you can automatically add new real-time and near-real-time GOES data into a queue or trigger event-based processing if the data meets certain criteria such as geographic location.

The ARN for the PDA feed is **arn:aws:sns:us-east-1:123901341784:NewGOES16Object.**

About the Data

Source	National Oceanic and Atmospheric Administration
Category	Earth Science, Sensor Data, Natural Resource, Meteorological
Format	netCDF v4
License	There are no restrictions on the use of this data.
Storage Service	Amazon S3
Location	s3://noaa-goes16 in us-east-1 region
Update Frequency	New data is added as soon as it's available

Earth on AWS Cloud Credits for Research

Educators, researchers and students can apply for free promotional credits to take advantage of Public Datasets on AWS. If you have a research project that could take advantage of GOES data on AWS, you can apply for Earth on AWS Cloud Credits for Research.

OCC's Environmental Data Commons

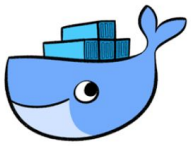
http://edc.occ-data.org/

The OCC Environmental Data Commons

Repository for environmental public data sets of scientific interest, hosted as part of the Open Science Data Cloud Ecosystem

GOES 16

NEXRAD



Tools | Notebooks

OCC GOES-16 Resources

http://edc.occ-data.org/goes16/

GOES-16 / GOES-R

Get Data

How to get GOES-16 data from the OCC Environmental Data Commons

Using Python to Explore GOES-16 Data

Working with GOES-16 data using Python and Jupyter

Manipulating GOES-16 Data with GDAL

Using GDAL to Work with GOES-16 NetCDF Data

OCC GOES-16 Resources

http://edc.occ-data.org/goes16/getdata/

Get Data

Provisional Data

NOAA's GOES-16 satellite has not been declared operational and its data are preliminary and undergoing testing.

Best Effort

The data are being made available on a best effort basis by all parties involved. There are no guarantees the data will be available when you really need it.