Enhancing Data Utilization Through Adoption of Cloud-Based Data Architecture

Jonathan S. O'Neil, Director, NOAA Big Data Project (BDP)
Jack Settelmaier, Technical Lead (NRAP)

National Oceanic and Atmospheric Administration

VLab Forum 8/21/2019

Outline

- What, Who, Why, How/Where
- Big Data Project - cadence, personnel, roles
- Data Usage Statistics / STI,R2O Use Cases
- Results/Successes/Issues/Future
- Questions

What

Our Mission

To understand and predict changes in climate, weather, oceans, and coasts, to ***share that knowledge and information*** with others, and to conserve and manage coastal and marine ecosystems and resources.

BDP Basics

- Cooperative Research and Development Agreements
  - 5 separate, but identical, 4-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
  - 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

Who

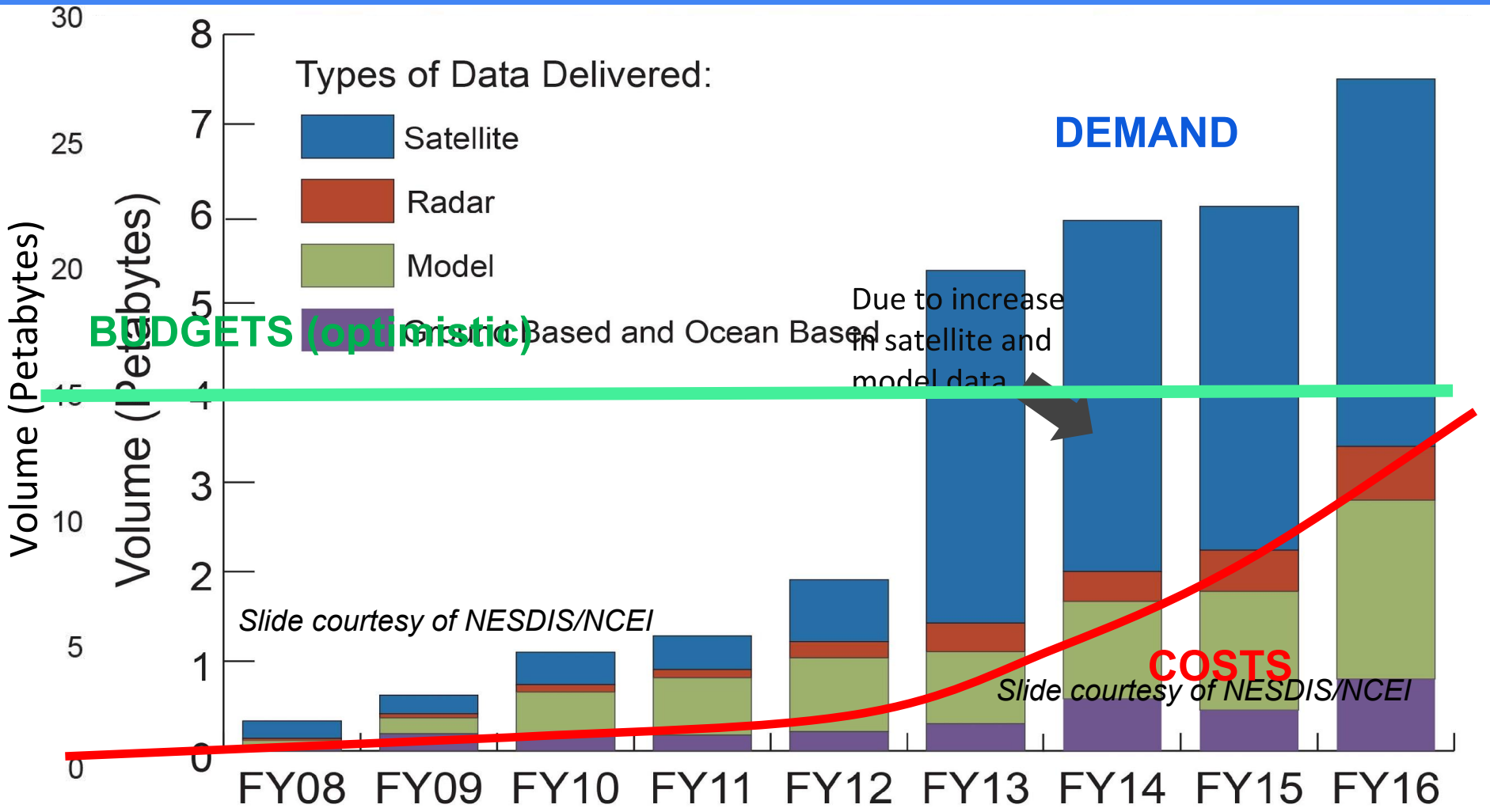
Big Data Project Collaborators' Data Offerings

- **AWS**
  - https://aws.amazon.com/noaa-big-data/
- **Google Cloud Platform**
  - https://cloud.google.com/bigquery/public-data/
  - https://explorer.earthengine.google.com/#index
- **IBM**
  - https://noaa-crada.mybluemix.net/
- **Microsoft**
  - https://azure.microsoft.com/en-us/services/open-datasets/catalog/nexrad-l2/
- **Open Commons Consortium**
  - http://edc.occ-data.org/

Why

Increasing Volume and Demand for NOAA Data

NOAA/NCEI's Environmental Data Archive



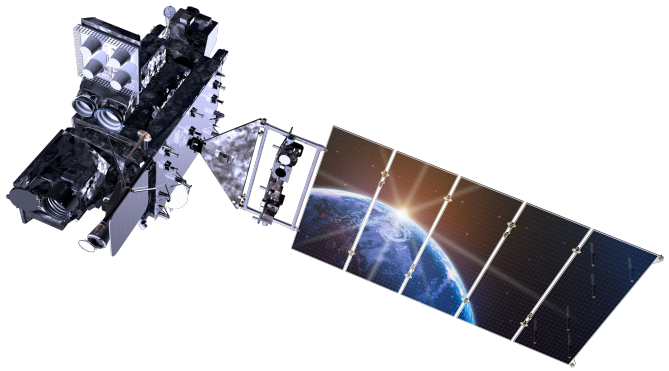
Collaborative Solutions

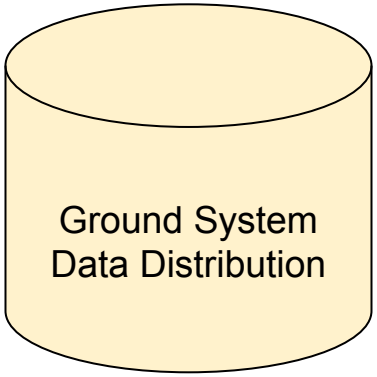
- Improve data access
- Facilitate use of the data
- Improve NOAA's cybersecurity posture
- Develop new authenticity tools
- Enable new economic & research opportunities

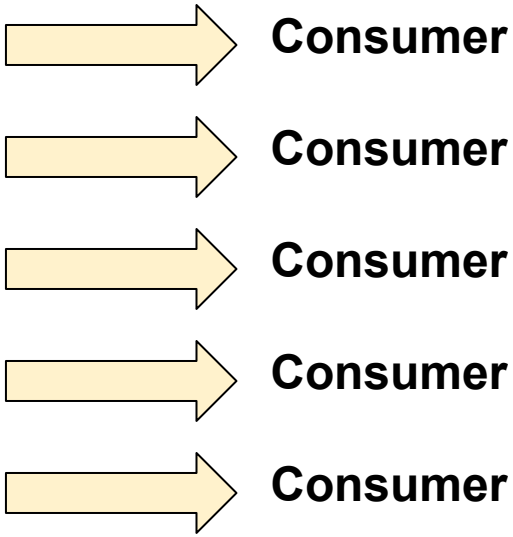
How/Where

Traditional NOAA Satellite Data Internet Access Strategy

Consumer Must Download Data to Use

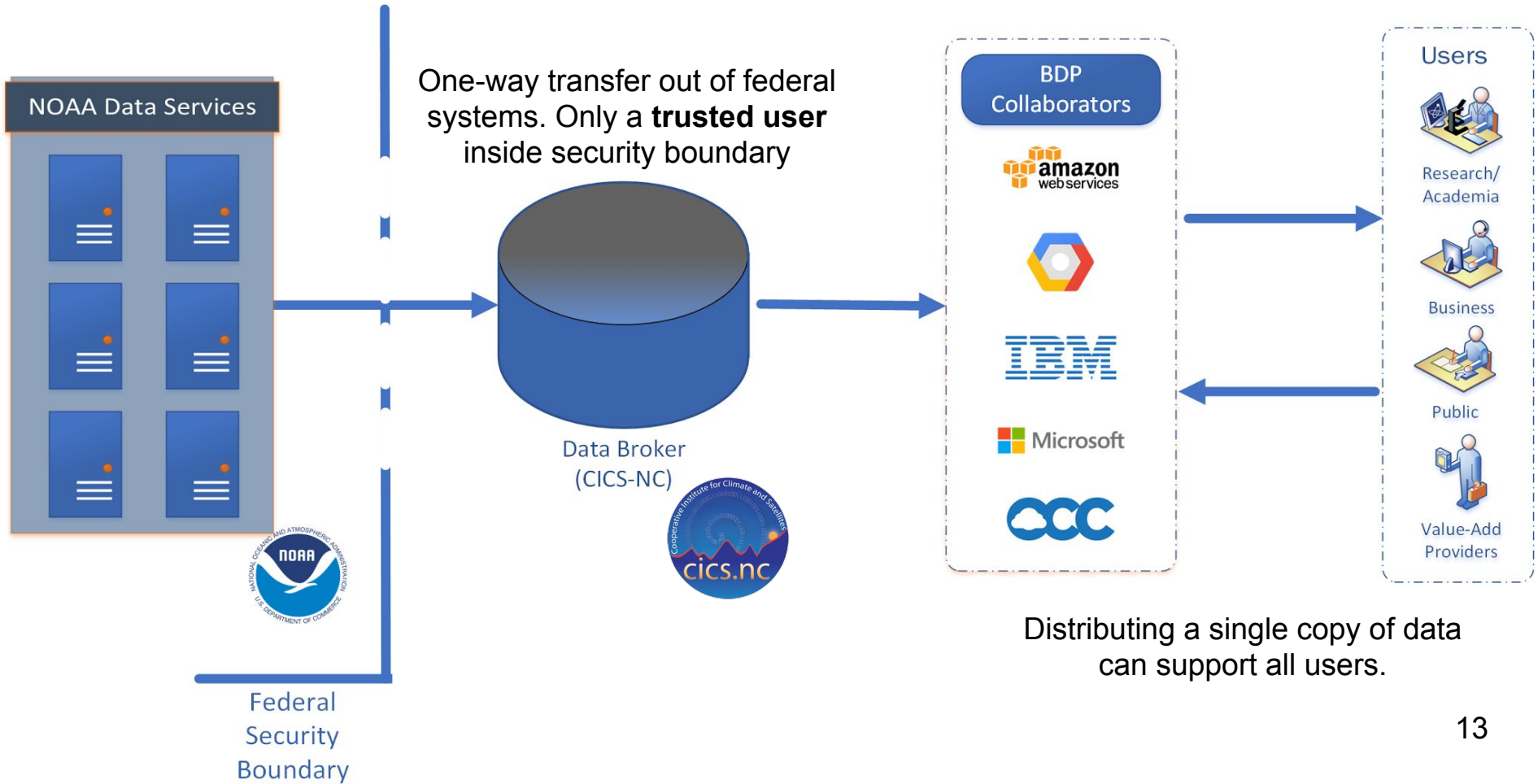






NOAA Big Data Project

Single Copy Services Many -- Use Data in Place



Distributing a single copy of data can support all users.

Data Broker Study: Format Conversions

- Pilot using Lambda functions converting GHCN-D granules from .gz to .csv:
  - Exceptions due to limitations built into Lambda functions (exceeds memory limit), requiring cloud instance to convert for most
- Datasets under discussion:
  - NetCDF images -> cloud optimized geotiff
  - GRIB/GRIB2 data matrices -> NetCDF or other formats

Dataset Maintenance

- NWC requested that we remove a parameter from the historical NWM datasets
  - Pilot using AWS FARGATE launching 20 *Docker* containers with
    - *Python* wrapper around the NCO (NCKS) tool
    - ~26 hours to remove one parameter from ~75K objects

The Limiting Factor for Use of the Data

Use

Cloud

Open
Data

Expertise

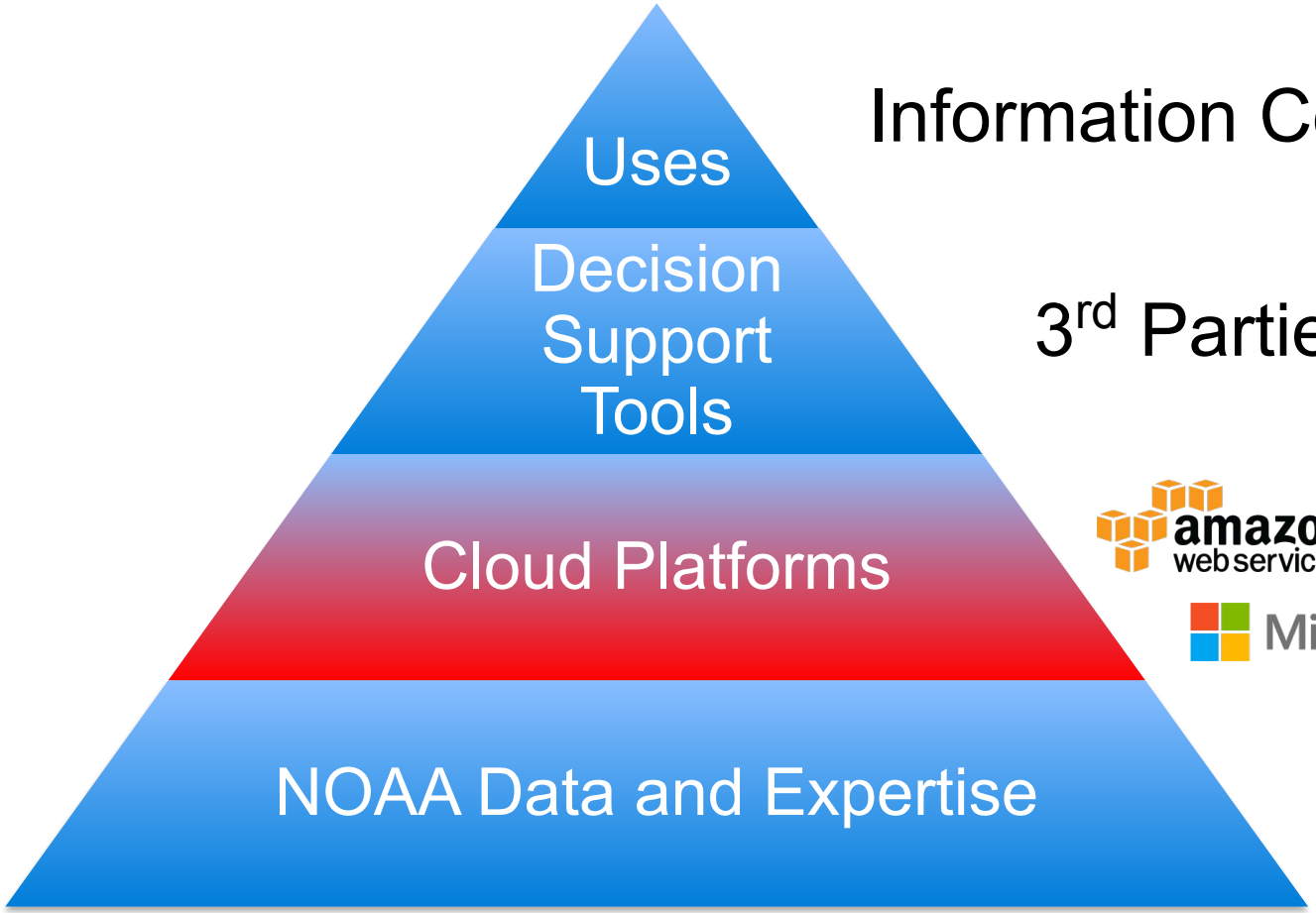
amazon web services

Microsoft

IBM

Google Cloud Platform

Value Driven Ecosystem



Information Consumers

3rd Parties

amazon web services

IBM

Google Cloud Platform

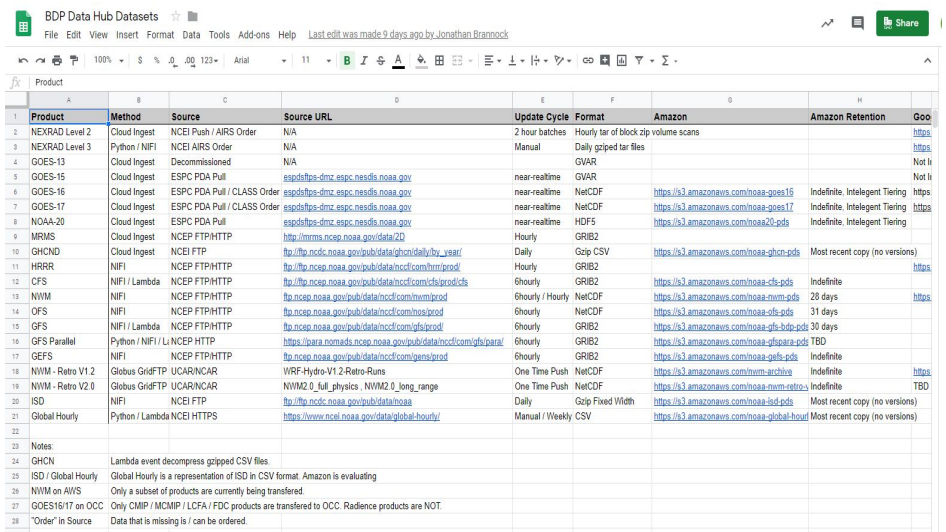
Microsoft

NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
U.S. DEPARTMENT OF COMMERCE

BDP Project Cadence Personnel, Roles

BDP Project Cadence Personnel, Roles

- Weekly/bi-weekly/monthly, per collaborator
- NOAA “data pitches” to collaborators
- Communications/Scheduling/Documenting
- Technical and non-technical NOAA liaison staff
- BDP Advisory Board



BDP Data Hub Datasets

File Edit View Insert Format Data Tools Add-ons Help Last edit was made 9 days ago by Jonathan Brannock

Share

Product

Product	Method	Source	Source URL	Update Cycle	Format	Amazon	Amazon Retention	Goo
NEXRAD Level 2	Cloud Ingest	NCEI Push / AIRS Order	N/A	2 hour batches	Hourly tar of block zip volume scans			https
NEXRAD Level 3	Python / NIFI	NCEI AIRS Order	N/A	Manual	Daily gziped tar files			https
GOES-13	Cloud Ingest	Decommissioned	N/A		GVAR			Not I
GOES-15	Cloud Ingest	ESPC PDA Pull	espdsftps-dmz.espc.nesdis.noaa.gov	near-realtime	GVAR			Not I
GOES-16	Cloud Ingest	ESPC PDA Pull / CLASS Order	espdsftps-dmz.espc.nesdis.noaa.gov	near-realtime	NetCDF	https://s3.amazonaws.com/noaa-goes16	Indefinite, Intelegent Tiering	https
GOES-17	Cloud Ingest	ESPC PDA Pull / CLASS Order	espdsftps-dmz.espc.nesdis.noaa.gov	near-realtime	NetCDF	https://s3.amazonaws.com/noaa-goes17	Indefinite, Intelegent Tiering	https
NOAA-20	Cloud Ingest	ESPC PDA Pull	espdsftps-dmz.espc.nesdis.noaa.gov	near-realtime	HDF5	https://s3.amazonaws.com/noaa20-pds	Indefinite, Intelegent Tiering	
MRMS	Cloud Ingest	NCEP FTP/HTTP	http://mrms.ncep.noaa.gov/data/2D	Hourly	GRIB2			
GHCND	Cloud Ingest	NCEI FTP	ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/	Daily	Gzip CSV	https://s3.amazonaws.com/noaa-ghcn-pds	Most recent copy (no versions)	
HRRR	NIFI	NCEP FTP/HTTP	ftp://ftp.ncep.noaa.gov/pub/data/nccf/com/hrrr/prod/	Hourly	GRIB2			https
CFS	NIFI / Lambda	NCEP FTP/HTTP	ftp://ftp.ncep.noaa.gov/pub/data/nccf/com/cfs/prod/cfs	6hourly	GRIB2	https://s3.amazonaws.com/noaa-cfs-pds	Indefinite	
NWM	NIFI	NCEP FTP/HTTP	ftp.ncep.noaa.gov/pub/data/nccf/com/nwm/prod	6hourly / Hourly	NetCDF	https://s3.amazonaws.com/noaa-nwm-pds	28 days	https
OFS	NIFI	NCEP FTP/HTTP	ftp.ncep.noaa.gov/pub/data/nccf/com/nos/prod	6hourly	NetCDF	https://s3.amazonaws.com/noaa-ofs-pds	31 days	
GFS	NIFI / Lambda	NCEP FTP/HTTP	ftp.ncep.noaa.gov/pub/data/nccf/com/gfs/prod/	6hourly	GRIB2	https://s3.amazonaws.com/noaa-gfs-bdp-pds	30 days	
GFS Parallel	Python / NIFI / La	NCEP HTTP	https://para.nomads.ncep.noaa.gov/pub/data/nccf/com/gfs/para/	6hourly	GRIB2	https://s3.amazonaws.com/noaa-gfspara-pds	TBD	
GEFS	NIFI	NCEP FTP/HTTP	ftp.ncep.noaa.gov/pub/data/nccf/com/gens/prod	6hourly	GRIB2	https://s3.amazonaws.com/noaa-gefs-pds	Indefinite	
NWM - Retro V1.2	Globus GridFTP	UCAR/NCAR	WRF-Hydro-V1.2-Retro-Runs	One Time Push	NetCDF	https://s3.amazonaws.com/nwm-archive	Indefinite	https
NWM - Retro V2.0	Globus GridFTP	UCAR/NCAR	NWM2.0_full_physics , NWM2.0_long_range	One Time Push	NetCDF	https://s3.amazonaws.com/noaa-nwm-retro-v	Indefinite	TBD
ISD	NIFI	NCEI FTP	ftp://ftp.ncdc.noaa.gov/pub/data/noaa	Daily	Gzip Fixed Width	https://s3.amazonaws.com/noaa-isd-pds	Most recent copy (no versions)	
Global Hourly	Python / Lambda	NCEI HTTPS	https://www.ncei.noaa.gov/data/global-hourly/	Manual / Weekly	CSV	https://s3.amazonaws.com/noaa-global-hourl	Most recent copy (no versions)	
Notes:								
GHCN	Lambda event decompress gzipped CSV files.							
ISD / Global Hourly	Global Hourly is a representation of ISD in CSV format. Amazon is evaluating							
NWM on AWS	Only a subset of products are currently being transfered.							
GOES16/17 on OCC	Only CMIP / MCMIP / LCFA / FDC products are transfered to OCC. Radience products are NOT.							
"Order" in Source	Data that is missing is / can be ordered.							

Data Usage Statistics
STI, R2O Use Cases

Big Data Project Collaborators' Data Offerings

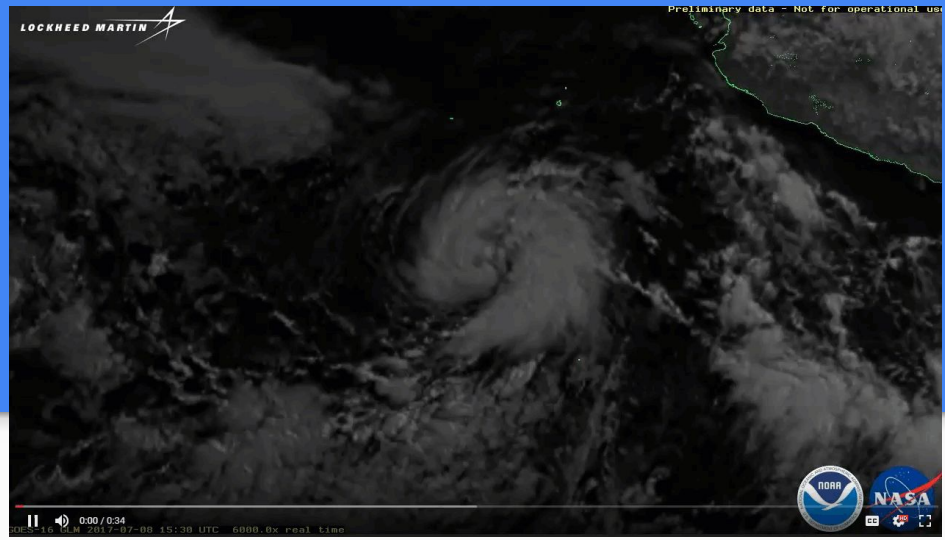
- **AWS**
  - https://aws.amazon.com/noaa-big-data/
- **Google Cloud Platform**
  - https://cloud.google.com/bigquery/public-data/
  - https://explorer.earthengine.google.com/#index
- **IBM**
  - https://noaa-crada.mybluemix.net/
- **Microsoft**
  - https://azure.microsoft.com/en-us/services/open-datasets/catalog/nexrad-l2/
- **Open Commons Consortium**
  - http://edc.occ-data.org/

NOAA's Big Data Project Use Cases

This document contains several use cases the Big Data Project (BDP) team has learned about from communications with end users and the Collaborators of the project. It is not until this last year of the experimental phase of the project that the BDP team has learned more about the various ways individuals, researchers, startups, and companies of different sizes have been utilizing the data accessed through the project. The BDP also gleaned valuable information from RFI responses. The RFI was published on October 1st 2018 on the Federal Register with an objective of informing the future direction of the Big Data Partnership. In that notice, NOAA indicated that it sought direct input and feedback from all users of the data on their experience accessing NOAA's open data through one or more of the five BDP Collaborators' cloud platforms.This document is a summary aimed at capturing the various use cases in greater detail, grouped by dataset.

References in this document to individuals, organizations, products, and services do not connote endorsement by NOAA or the Department of Commerce of such individuals, organizations, products, or services, or of their views.

BDP and GOES-16



Overview of BDP Activity

Cooperative Institute for Climate and Satellites – North Carolina

- Providing GOES-16 data from NOAA/NESDIS Ground System (PDA) to Collaborators.

1 source, 5 validated feeds to the CRADA Collaborators

- Timing - As fast as they appear at NOAA/NESDIS distribution point
- Latency - Single hop through CICS-NC systems, w/checksums
- Impact - Minimal load on NOAA's operational systems and networks

Observed additional latencies from CICS-NC transfer 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: 4.5 to 5.5 seconds**

AWS and Weather Radar

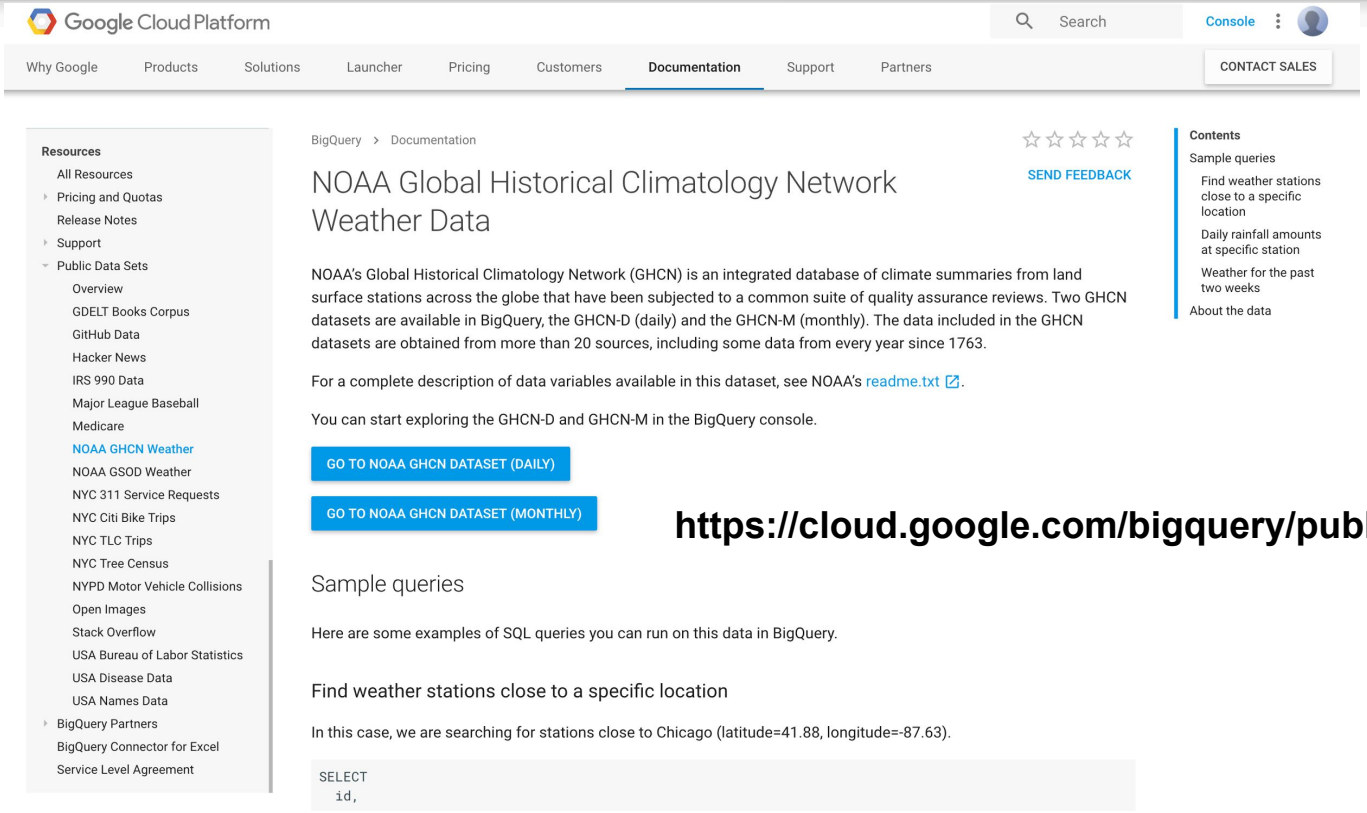
Entire NEXRAD 88D Weather Radar Archive transferred to AWS, Google and OCC in Oct 2015 (~ 300TB, 20M files)

Following AWS service release:

- Increased usage (2.3 times), 50% reduction on NOAA servers.
- New uses – bird migration, mayfly studies

- 80% of NOAA NEXRAD data orders are now served by AWS.
  - *(Ansari et al, 2017 BAMS)*

Climate Data in Google BigQuery



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

- **1.2 PBs** of climate and weather data accessed through Google BigQuery, in **4 months**
  - 30-100x of NOAA deliveries in that time
- Images in Google Cloud Platform
  - GOES-16 (began July 2017)
  - National Water Model data
  - Weather and Climate model output
  - Climate data records

OCC's Environmental Data Commons

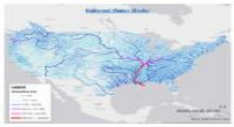
http://edc.occ-data.org/

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



NWM

Tools | Notebooks

Tampa Bay Times

A matter of miles

How the slightest shift kept Hurricane Irma from turning into an even worse disaster

By NATHANIEL LASH and NEIL BEDI

Times Staff Writers

Sept. 20, 2017

Scroll down

↓

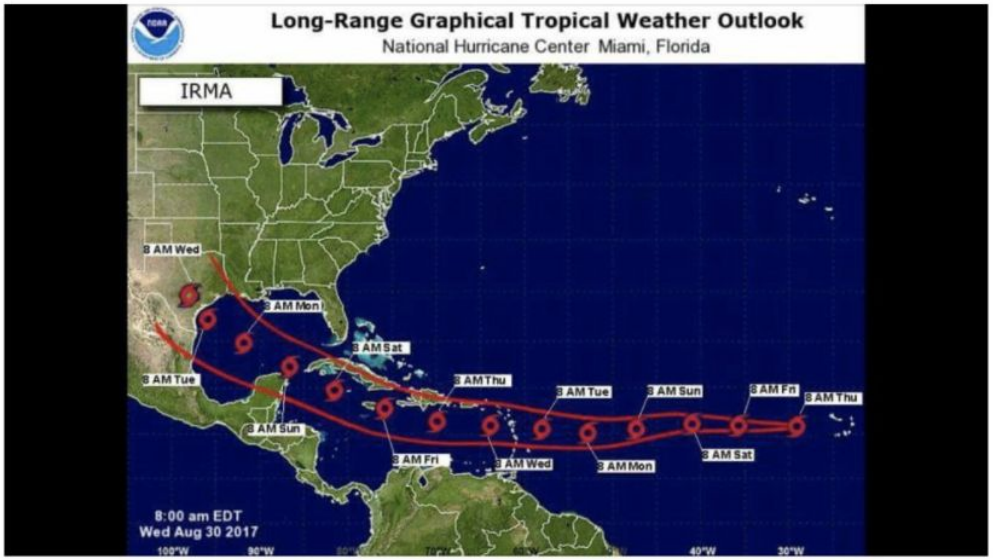
Ensuring Data Authenticity

- Users trust NOAA data they access from NOAA sites
- What about outside NOAA?
  - Collaborators' sites?
- Example: Fake Irma forecast

Nearly 40,000 shares on this fake Irma forecast on Facebook. There is no actual threat to any land in next 5 days.
pic.twitter.com/hwmuE5UwKn
10:11 AM - Sep 1, 2017

[redacted]

Everyone needs to pay attention to Hurricane Irma. Shes predicted to come through mexico hit us and everything inbetween up to Houston. Shes already a Category 2 and hasnt even got into warm water yet.



Like Comment Share

2.5K

36,955 shares

17 70 57

Data Broker's Usage Statistics Analysis (presented at Earth Science Information Partners (ESIP) meeting

CISESS
Cooperative Institute for
Satellite Earth System Studies

BDP Data Broker Update

Jonathan Brannock and Otis Brown
CISESS / NCICS
NCSU

NC STATE UNIVERSITY

Broker Roles

- Impedance matching with each side of the transfer, *i.e.,* use appropriate protocols for each interaction
- Transfer data from NOAA to Cloud partners
- Resolve operational problems
- Ensure data integrity
- Certify Cloud partner holdings
- Add new datasets

Broker Operations

- Inform Cloud partners of source data outages
- Resolve source and/or destination outages
- Resolve transmission errors
- Maintain 24/7 operations with minimal data loss
- Add new data sets as requested

Currently Brokered Datasets – ~5TB / day

Dataset	Size	Cloud Provider(s)	Source
GOES-16/17 ABI & GLM	~1 TB daily	AWS, GCS, OCC	NESDIS/ESPC
GOES-15 GVAR	~15 GB daily	OCC	NESDIS/ESPC
NEXRAD L2	~60 - 201 GB daily	AWS, GCS, OCC	NESDIS/NCEI
NEXRAD L3	~20 - 60 GB daily	GCS	NESDIS/NCEI
GHCN-D	~30 GB daily	AWS	NESDIS/NCEI
National Water Model	~115 GB daily	AWS, GCS	NWS/NCO
Ocean Forecast System	~288 GB daily	AWS	NWS/NCO

Brokered Datasets (Continued)

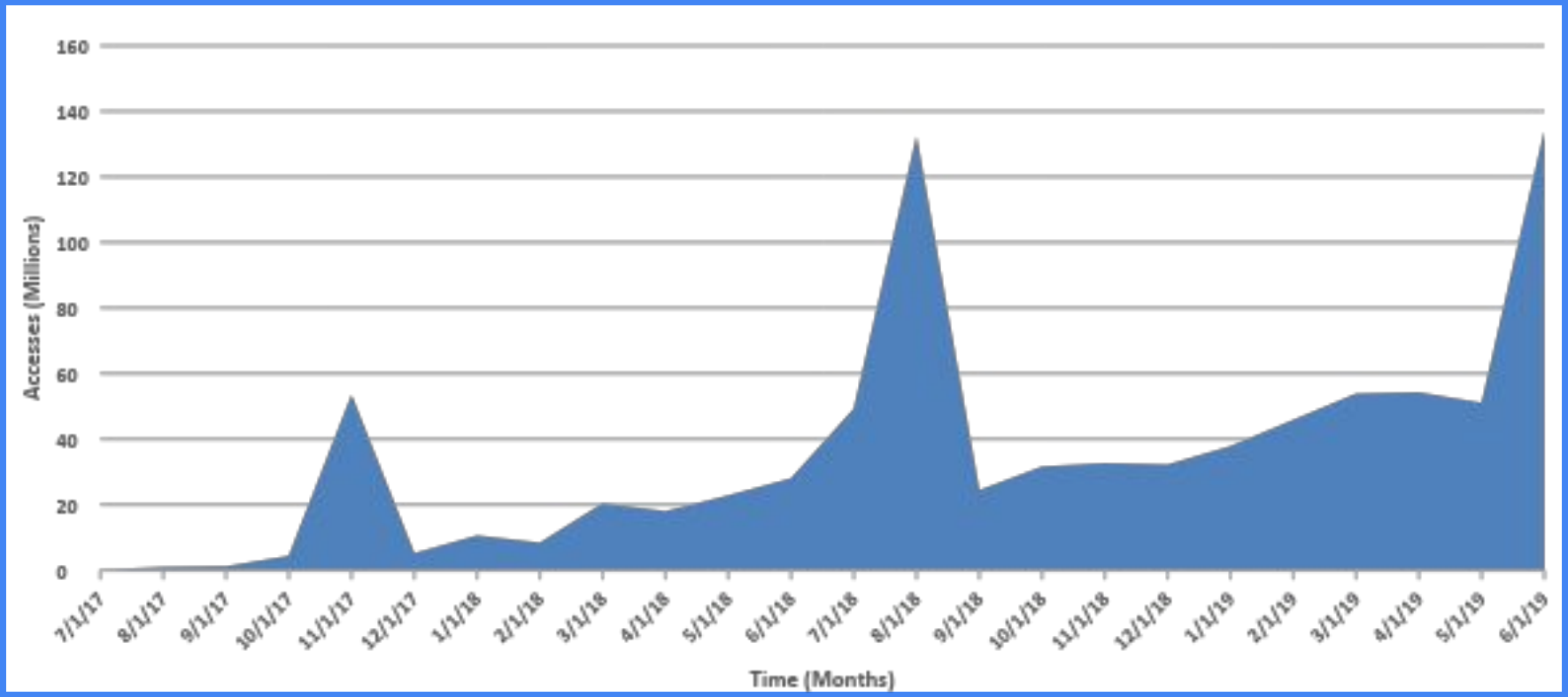
Dataset	Size	Cloud Provider(s)	Source
NOAA-20 VIIRS DNB	~60 GB daily	AWS	NESDIS/ESPC
Fire Products	~1.1 GB daily	AWS,GCS	NESDIS/ESPC
ISD	~750 MB daily	AWS	NESDIS/NCEI
Global Hourly	~20 GB daily	AWS	NESDIS/NCEI
GFS	~950 GB daily	AWS	NOS/CO-OPS
HRRR	~920 GB daily	AWS,GCS	NWS/NCO
CFS	~310 GB daily	AWS	NWS/NCO

Brokered Datasets Accession (Ranked)

Dataset	Count	Dataset	Bytes
GOES-16	754,659,402	GOES-16	3,163,532GB
GOES-17	80,221,671	GOES-17	584,946GB
NWM (Archive)	11,530,076	NWM (Archive)	142,768GB
GFS	3,048,159	GHCNd	50,658GB
GHCNd	1,116,821	GFS	41.800GB
NWM	136,287	GEFS	.377GB
ISD	114,879	GFS(para)	.297GB
GEFS	63,598	OFS	.066GB

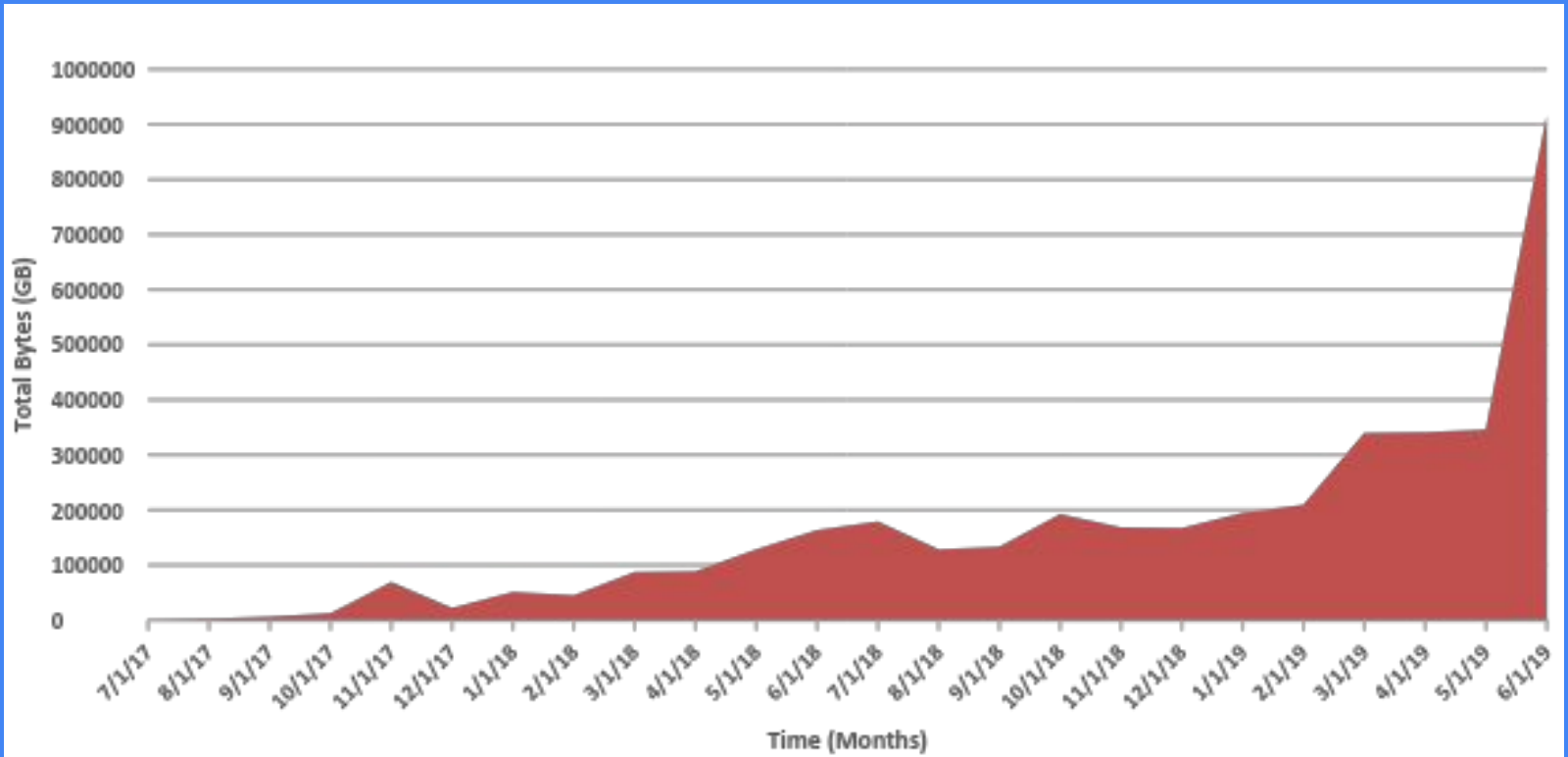
Statistics are for the total BDP period of record for each data set

Accession Rates



Statistics are for the BDP Brokered datasets

Accession Rates - Volume

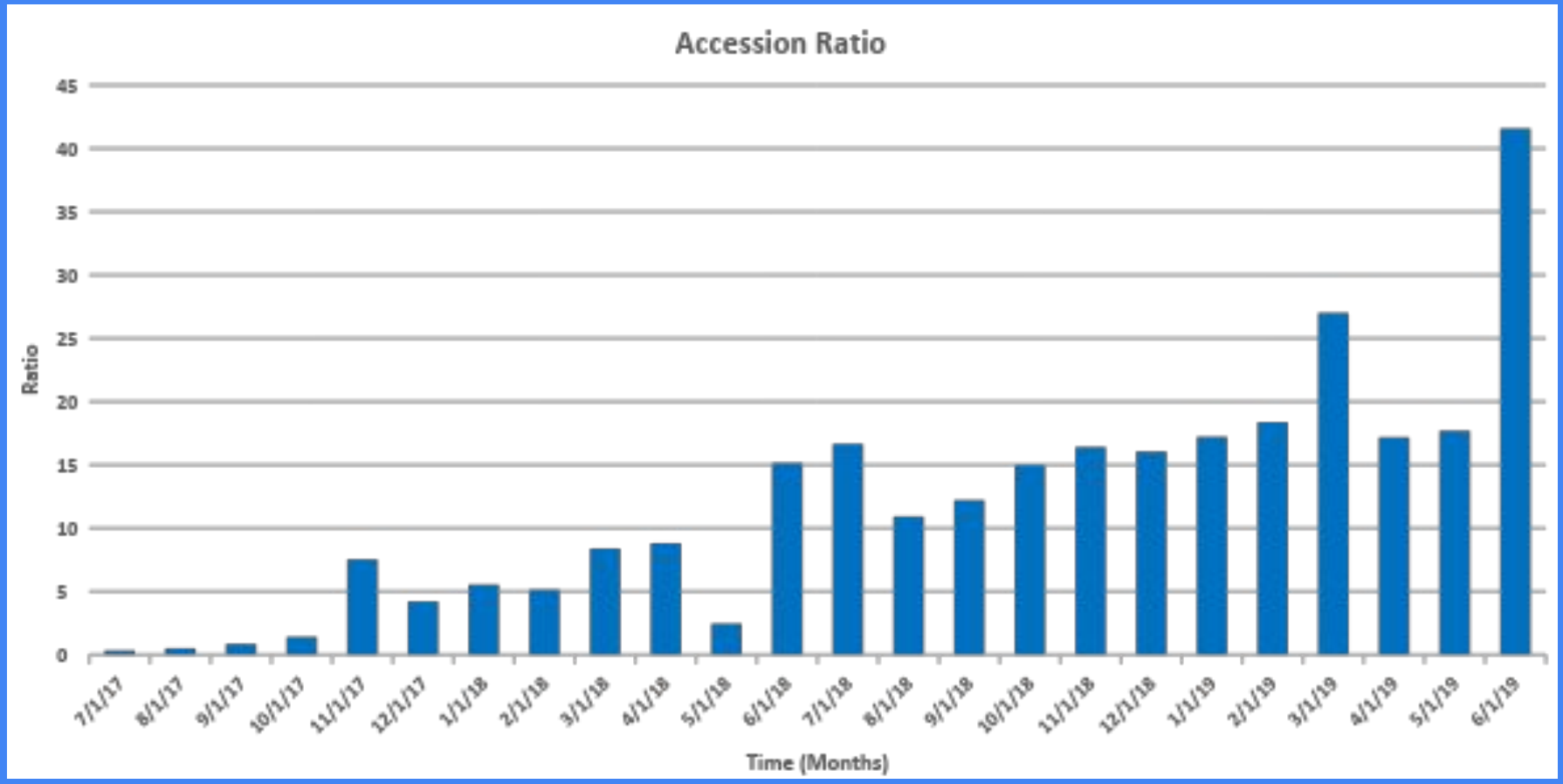


Trend is ~16.33 TB/Month increase (7/17 – 5/19)
~900 TB of GOES-16/17 accessed in 6/19

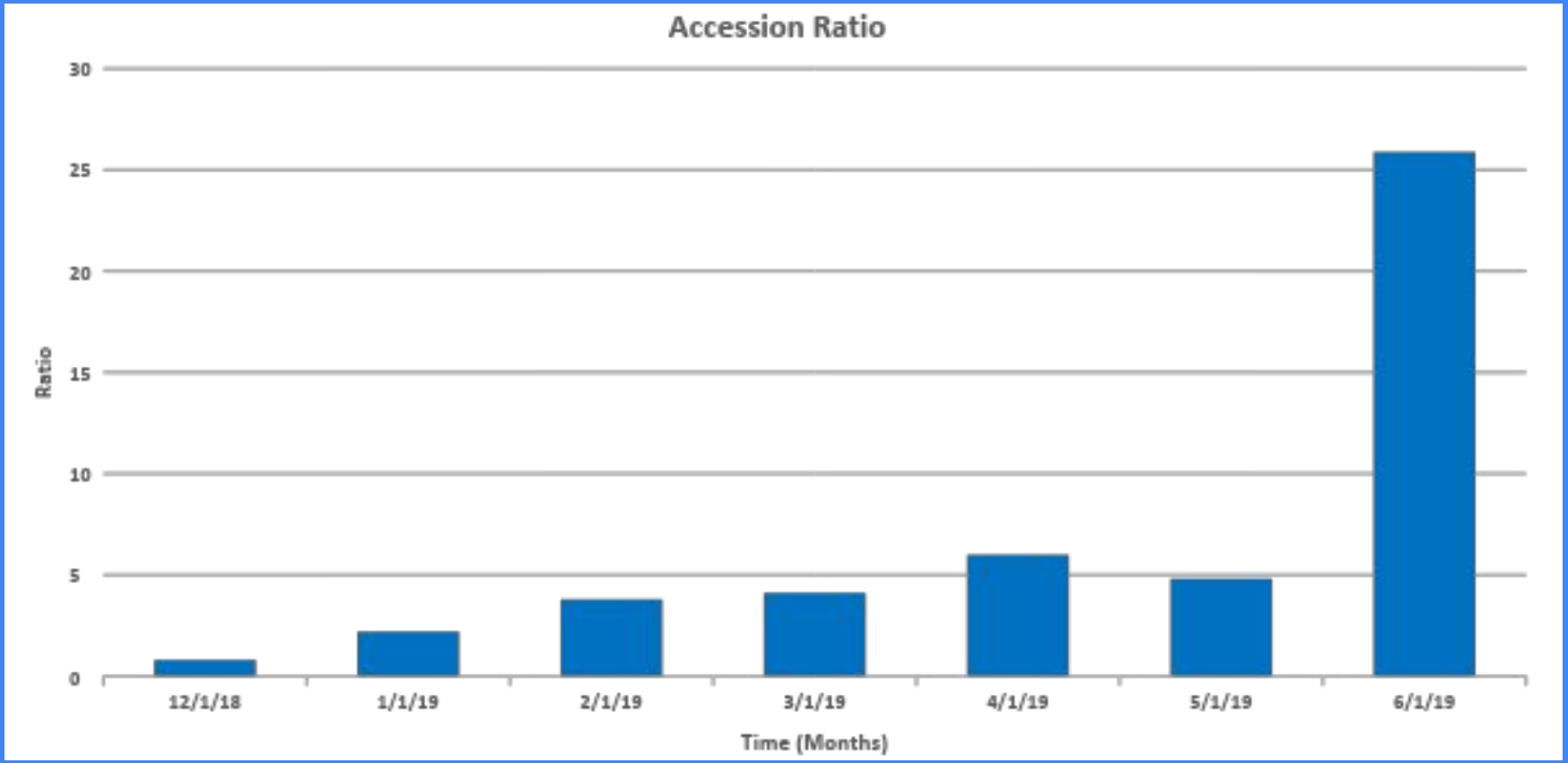
BDP Terminology

- BDP is “Big Data”
  - Volume – TB per day
  - Variety – NOAA environmental observations and model products
  - Velocity – Minimal latencies, in near-real time (secs – mins)

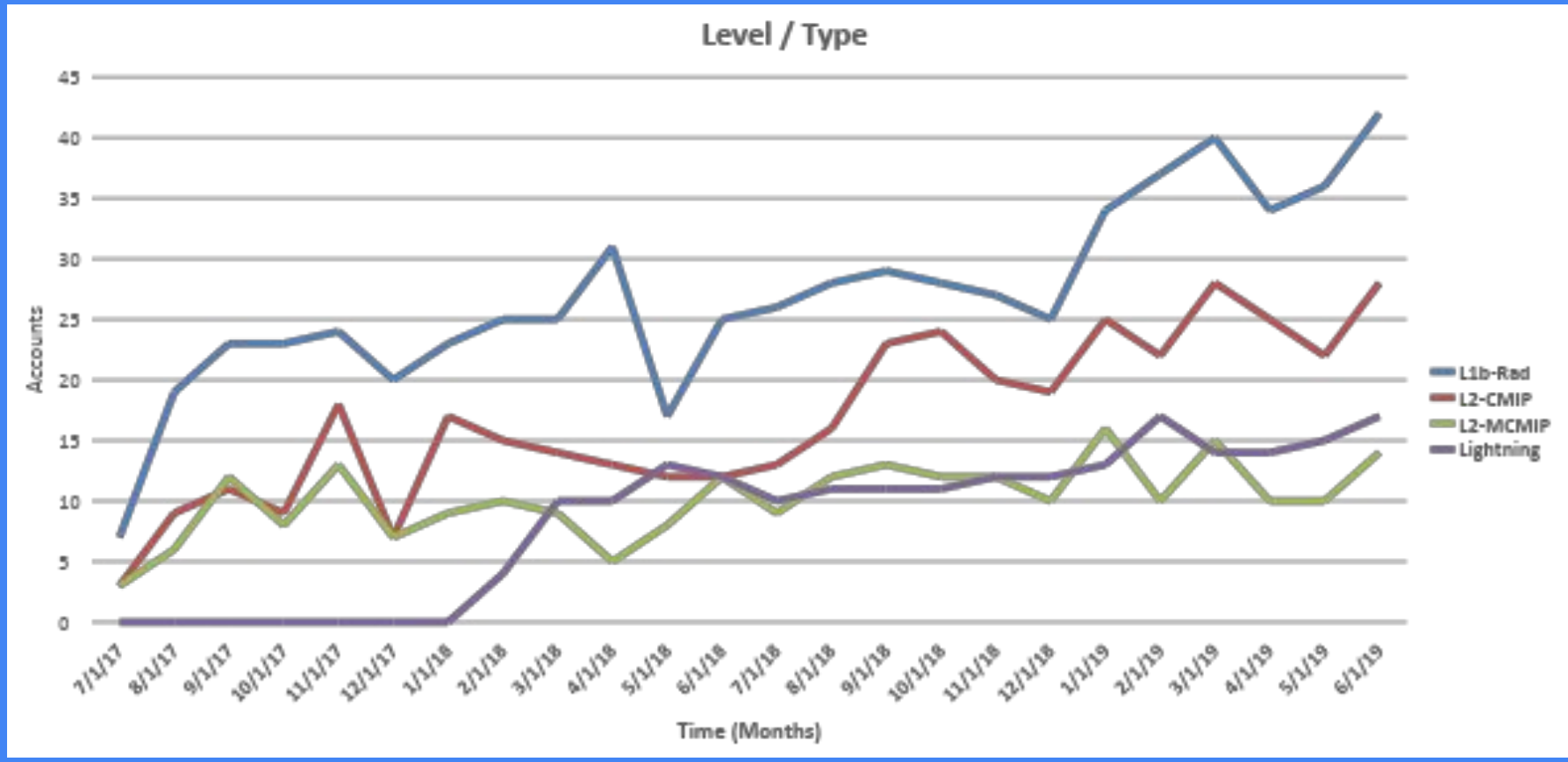
GOES-16 Data Statistics [AWS]



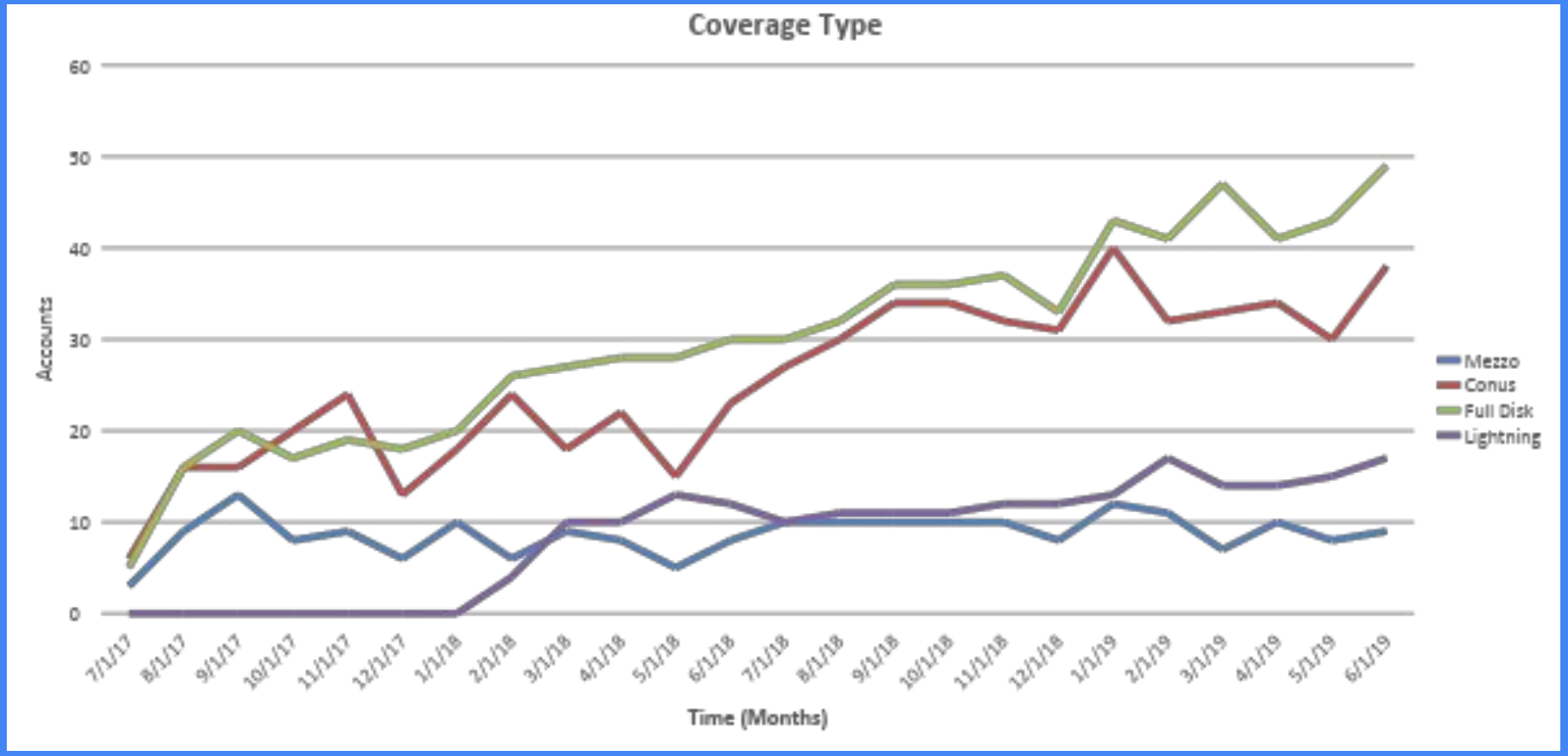
GOES-17 Data Statistics [AWS]



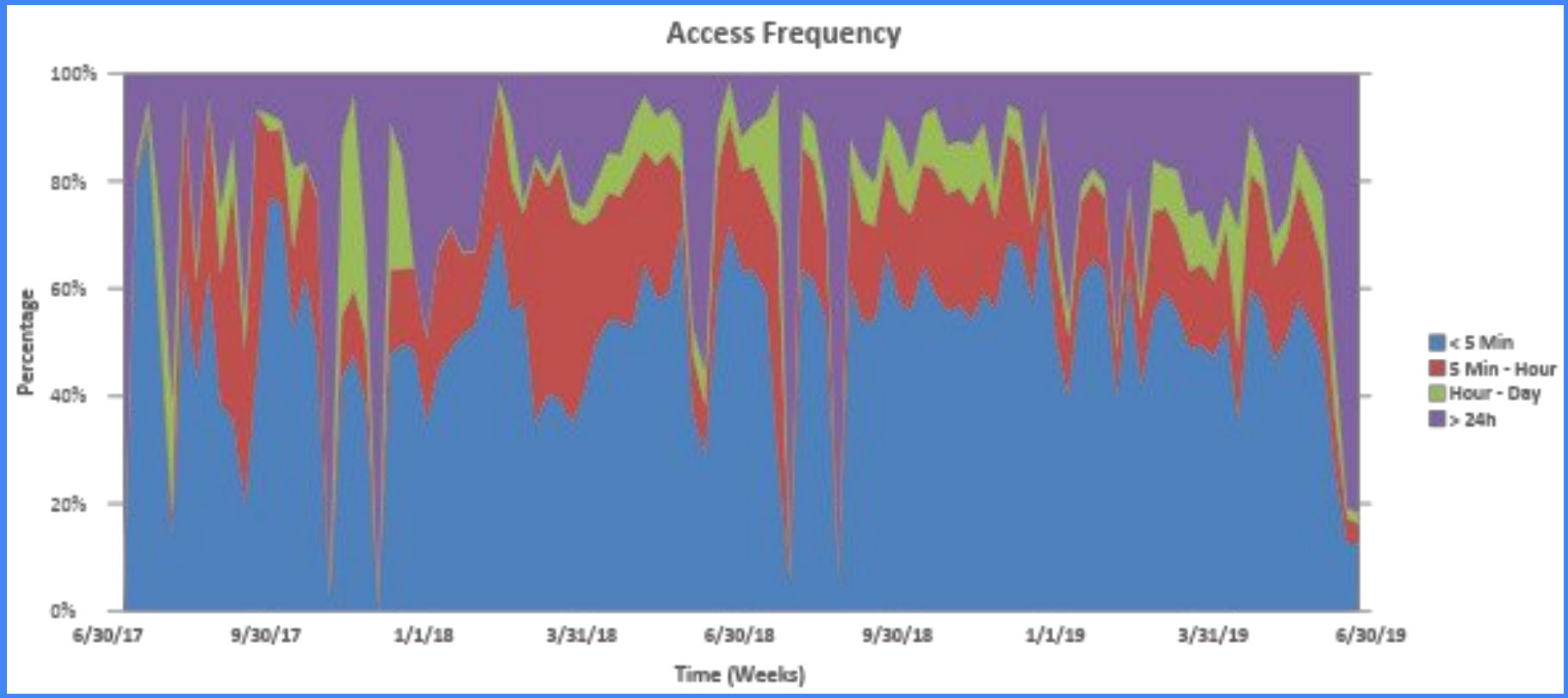
GOES-16/17 Data Statistics [AWS]



GOES-16/17 Data Statistics [AWS]

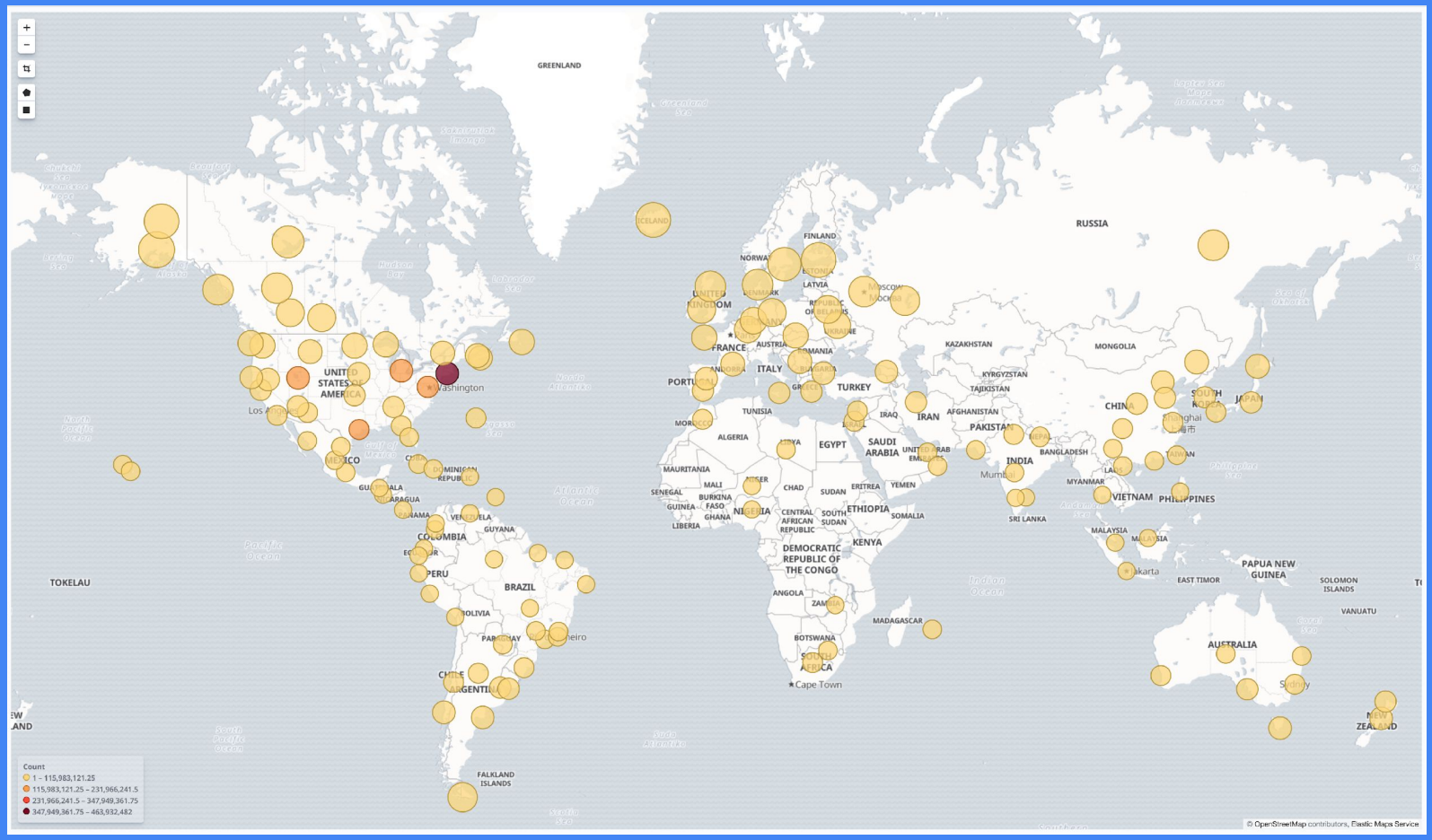


GOES-16/17 Access Statistics [AWS]

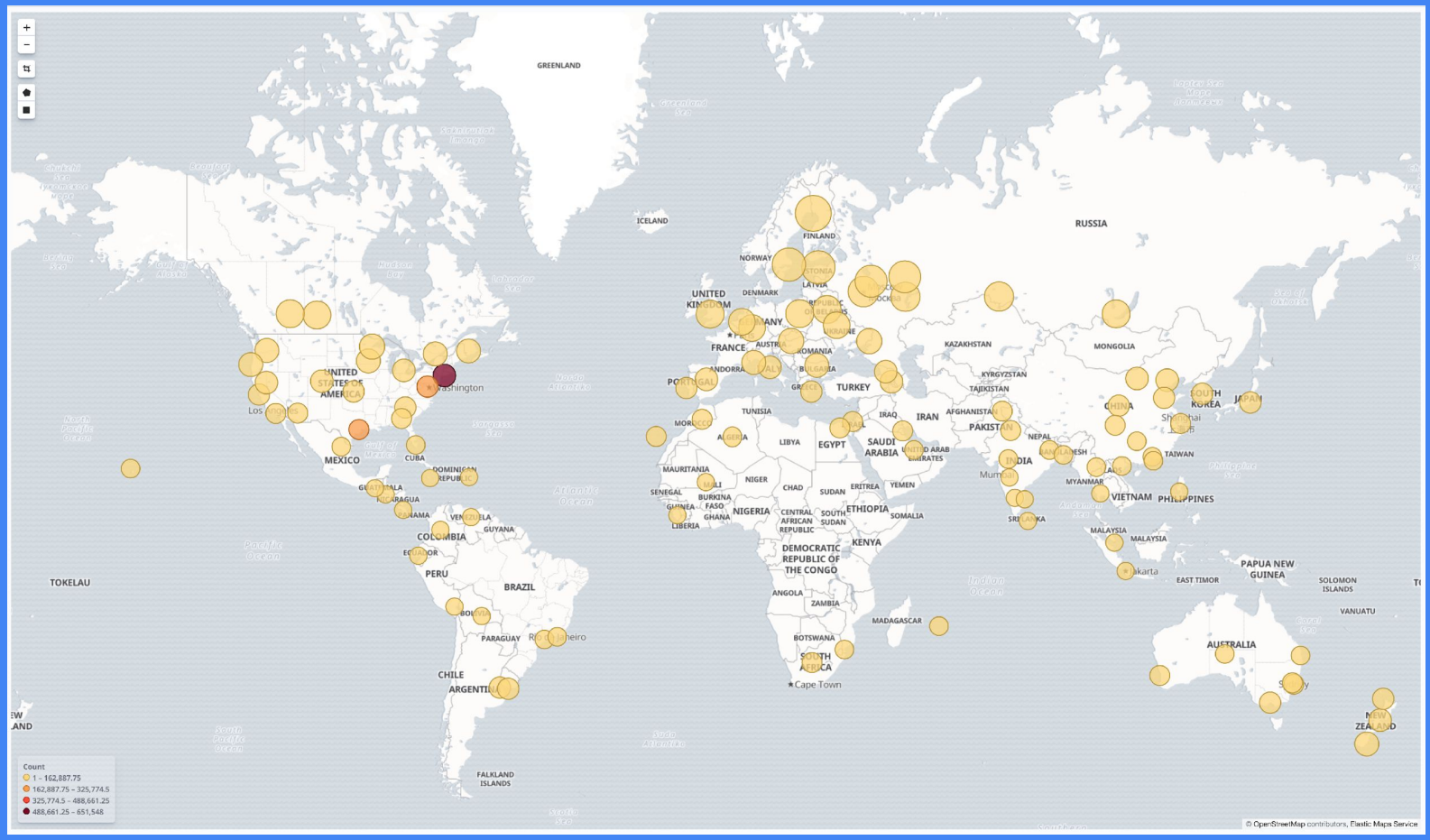


Age is the time from observation. Average frequencies are: ~50%, ~18%, ~8% and ~24%, respectively.

GOES-16 Access Locations [AWS]

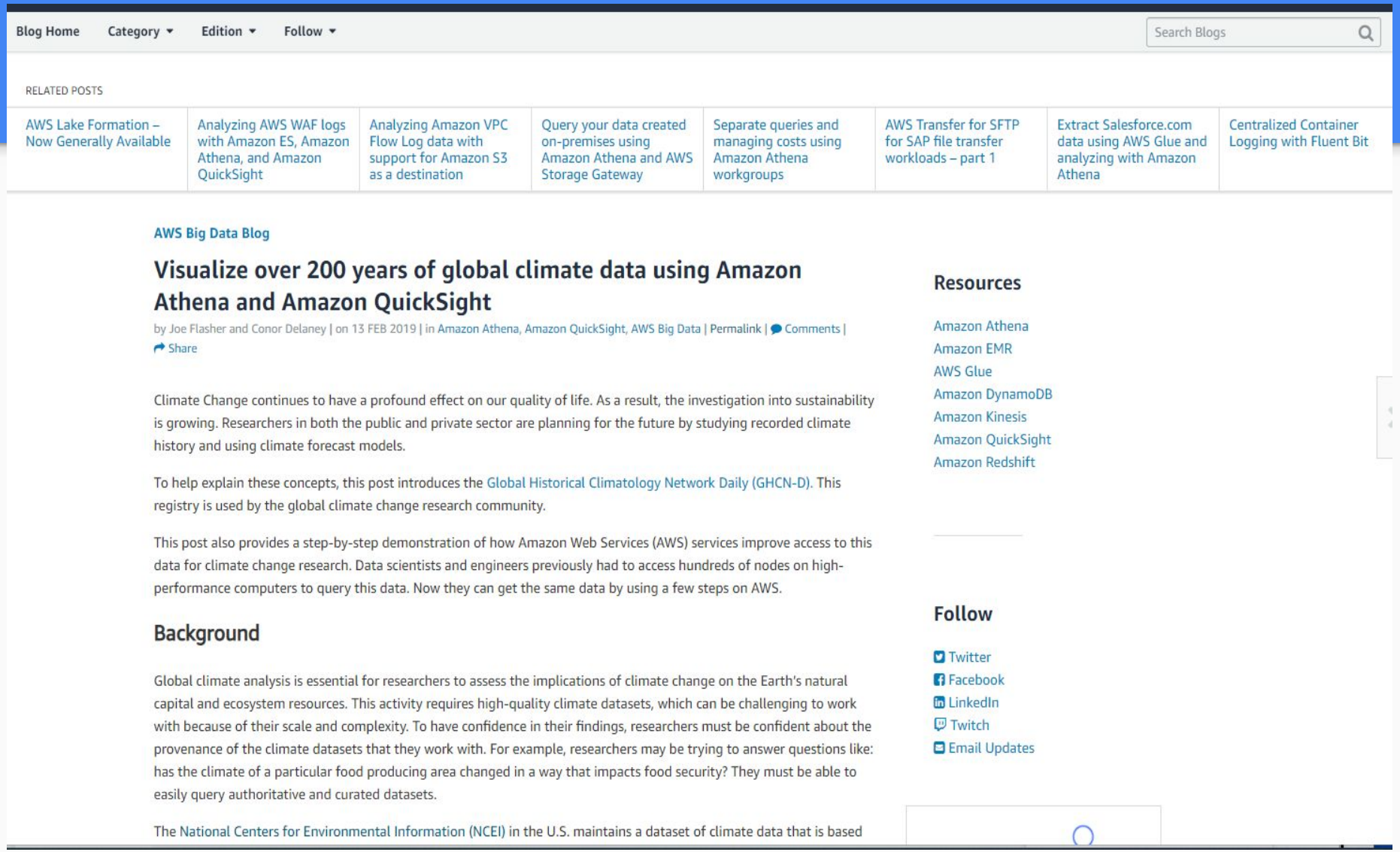


GHCNd Access Locations [AWS]



Data on Cloud Platforms
STI/R2O Use Cases

AWS GHCN-D Blog Post



Blog Home | Category | Edition | Follow | Search Blogs

RELATED POSTS

- AWS Lake Formation – Now Generally Available
- Analyzing AWS WAF logs with Amazon ES, Amazon Athena, and Amazon QuickSight
- Analyzing Amazon VPC Flow Log data with support for Amazon S3 as a destination
- Query your data created on-premises using Amazon Athena and AWS Storage Gateway
- Separate queries and managing costs using Amazon Athena workgroups
- AWS Transfer for SFTP for SAP file transfer workloads – part 1
- Extract Salesforce.com data using AWS Glue and analyzing with Amazon Athena
- Centralized Container Logging with Fluent Bit

AWS Big Data Blog

Visualize over 200 years of global climate data using Amazon Athena and Amazon QuickSight

by Joe Flasher and Conor Delaney | on 13 FEB 2019 | in Amazon Athena, Amazon QuickSight, AWS Big Data | Permalink | Comments | Share

Climate Change continues to have a profound effect on our quality of life. As a result, the investigation into sustainability is growing. Researchers in both the public and private sector are planning for the future by studying recorded climate history and using climate forecast models.

To help explain these concepts, this post introduces the Global Historical Climatology Network Daily (GHCN-D). This registry is used by the global climate change research community.

This post also provides a step-by-step demonstration of how Amazon Web Services (AWS) services improve access to this data for climate change research. Data scientists and engineers previously had to access hundreds of nodes on high-performance computers to query this data. Now they can get the same data by using a few steps on AWS.

Background

Global climate analysis is essential for researchers to assess the implications of climate change on the Earth's natural capital and ecosystem resources. This activity requires high-quality climate datasets, which can be challenging to work with because of their scale and complexity. To have confidence in their findings, researchers must be confident about the provenance of the climate datasets that they work with. For example, researchers may be trying to answer questions like: has the climate of a particular food producing area changed in a way that impacts food security? They must be able to easily query authoritative and curated datasets.

The National Centers for Environmental Information (NCEI) in the U.S. maintains a dataset of climate data that is based

Resources

- Amazon Athena
- Amazon EMR
- AWS Glue
- Amazon DynamoDB
- Amazon Kinesis
- Amazon QuickSight
- Amazon Redshift

Follow

- Twitter
- Facebook
- LinkedIn
- Twitch
- Email Updates

ESIP Presentation

Jupyter … Notebooks, lab, hub, collaboratory

Let's jump in! Head to…

Github.com

Search for this repository:

ExploreAtlanticStorms

Results/Successes/Issues
Future?

Big Data Project Results

Technical Proof-of-Concept proven successful

- Over 70 datasets are being served now via Collaborator services
- Higher levels of service, increased volumes, new users
- Lessons-learned documented; Collecting User stories

Key Points

- A Partnership based on NOAA public data access can work
- Expertise is the scarce, valuable commodity in the relationship
- NOAA must own data security & quality to secure value & brand

Opportunities that the Partnership aspect may provide

- Grow & defend NOAA budgets with new advocates for NOAA data
- Cost savings in data access infrastructure
- Enhanced scale and reliability of services for NOAA data consumers

Big Data Project
Detailed Successes

- **Over 70 datasets** are being served now via Collaborators (weather radar data, historical weather data, GOES 16/17 satellite imagery, lightning observations (GLM and Vaisala aggregate), fisheries data, and a variety of computer model outputs including the NWM, GFS, CFSv2 and HRRR.
- Collaborators **interested in continuing** partnership contractually, post-CRADA
- **Significant** increases in data usage have been observed*
  - GOES accession rates 10-15x + the incoming data rate
  - 130% increase in weather radar data use over previous years
  - 50% reduction in access loads on the NOAA systems
  - 80% of archive data orders now fulfilled on collaborators' systems.
- Collaborators have stated that **access to NOAA's expertise** has been the most valuable
- **Integration** of NOAA data into collaborators' **existing cloud-based access and analytical tools** has driven the largest increases in data usage.
- Activities/labor costs associated with data delivery from NOAA to cloud have been collapsed into a **"data broker"** role (NOAA's Cooperative Institute (CICS/CISESS))--identified as a KEY function for the project.

Big Data Project
Issues/Next Steps/Future

- Sustained operational phase?
- Cloud services strategy ties, cyber-security (data broker)
- Data authenticity (origin/quality); Data management (curation)
- User profiles and usage statistics
- Open and non-open data
- Budget and Cost Recovery (dedicated, not detailees)

- NOAA Implementation Strategy?
  - Request for Proposals issued April-May, 2019

Questions?


TIME WELL SPENT™
by Tom Fishburne
SEE, I TOLD YOU THAT BIG DATA WAS TOO SCARY
BIG DATA
KRONOS
Workforce Innovation That Works™
© 2012
KRONOS.COM/TIMEWELLSPENT

NOAA is seeking a Sustainable Partnership

- Developed a Conceptual and Functional Framework
  - Achieved Line Office Concurrence
- Defining Implementation Options
- Analyzing the User information available
- Gathering Use Cases
- **Request For Information conducted October 2018**
  - **Responses due October 22, 2018**
- **Request for Proposals issued April-May, 2019**