Data Dissemination Best Practices and Challenges Identified Through NOAA’s Big Data Project

Meredith Richardson, Jonathan O’Neil, Ed Kearns
NOAA Data Program, NOAA OCIO
National Oceanic and Atmospheric Administration

EGU May 2020

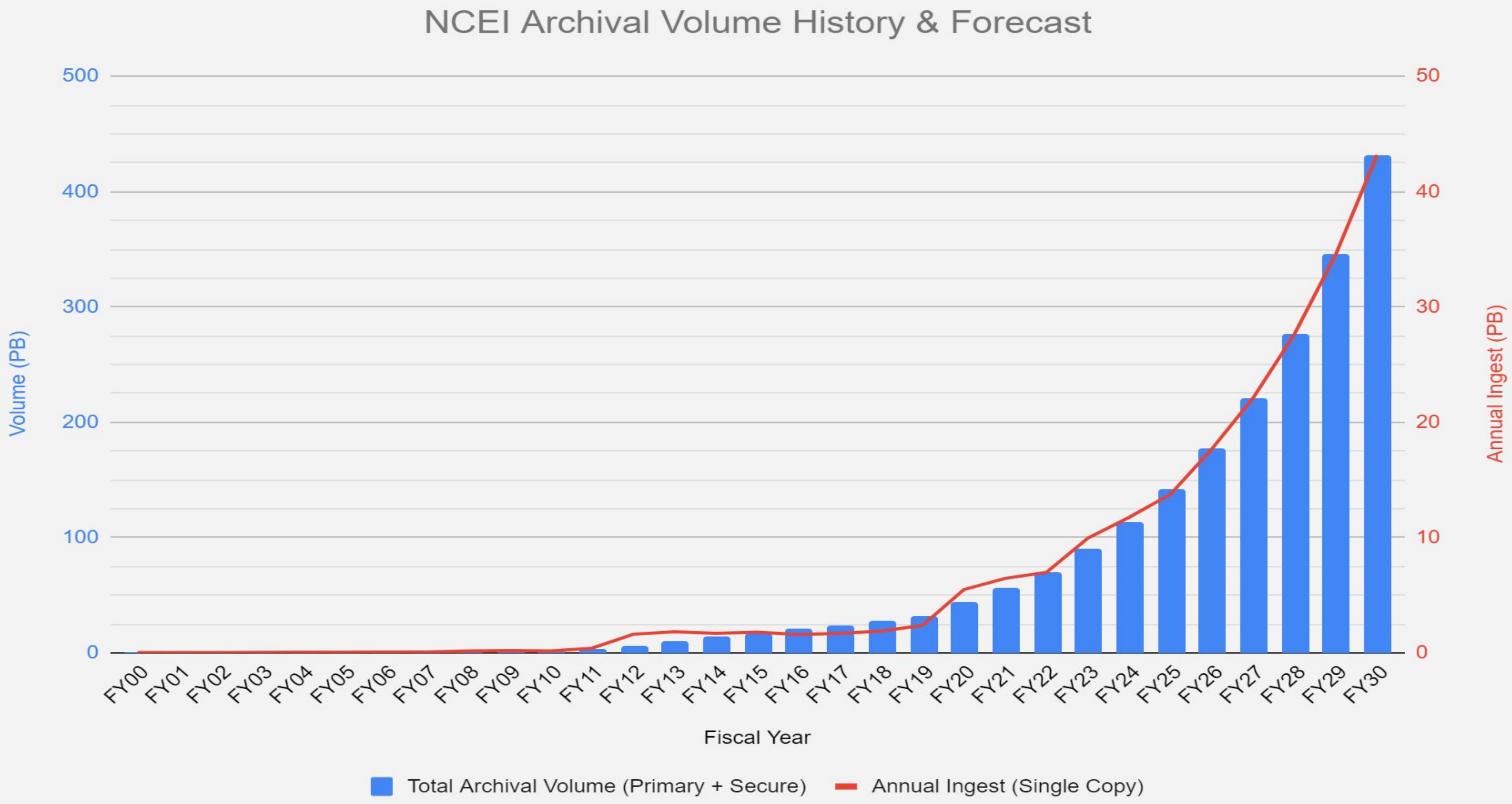
The Big Ideas

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.

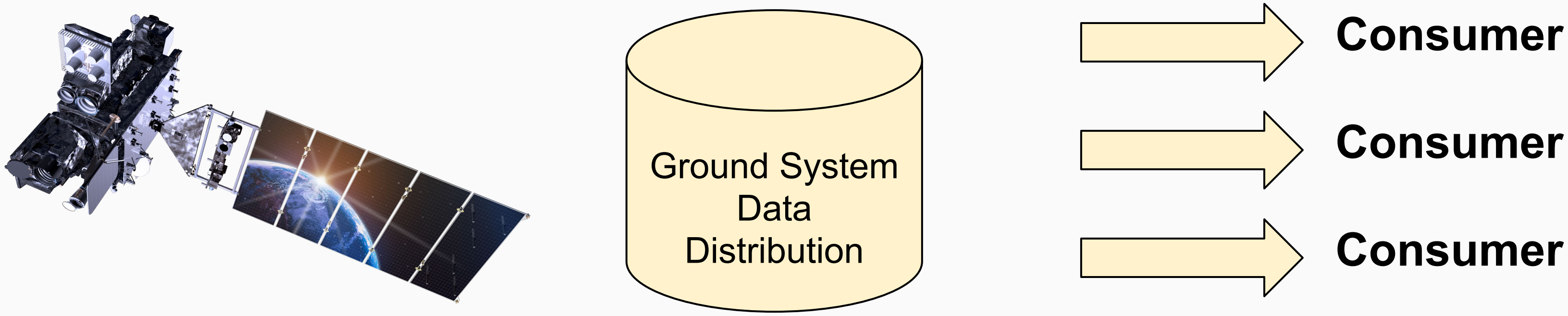
Motivation: Increasing Volume and Demand for NOAA Data

NOAA/NCEI's Environmental Data Archive



Traditional NOAA Satellite Data Internet Access Strategy

Challenge: Consumer Must Download Data to Use



Ground System Data Distribution

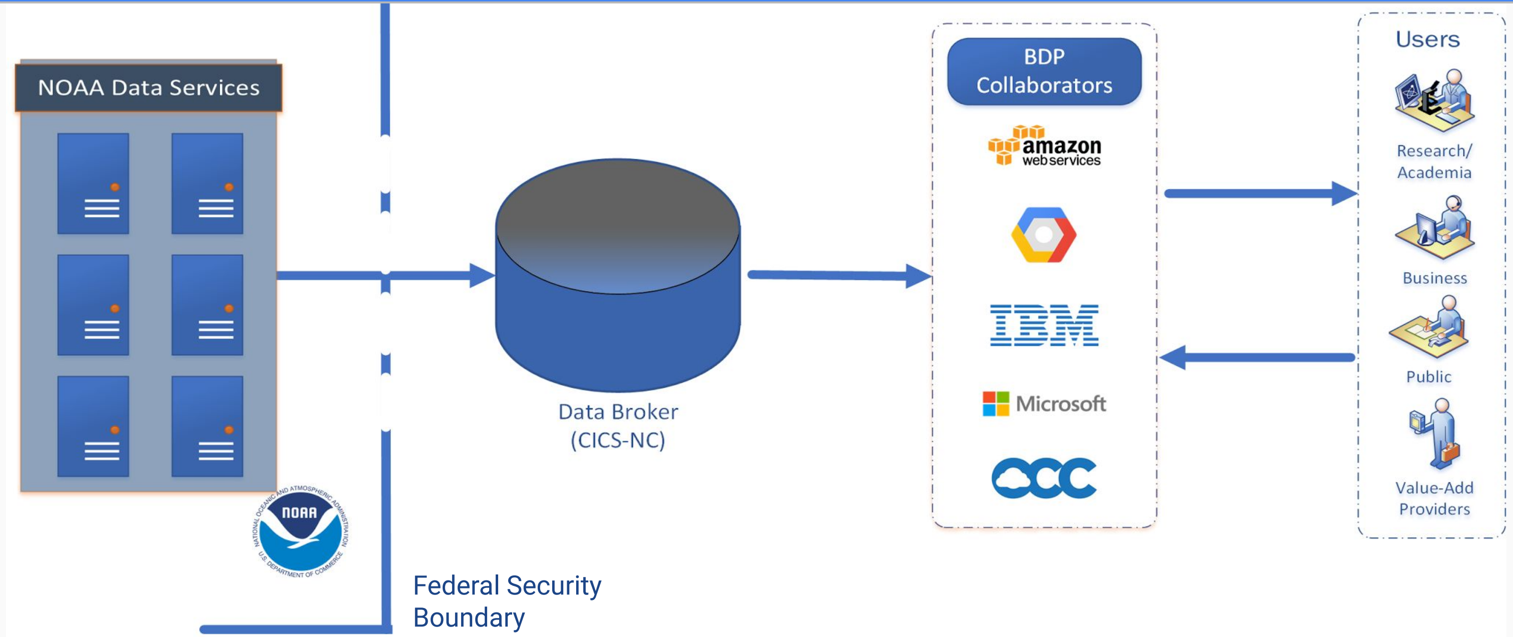
Consumer

Consumer

Consumer

NOAA Big Data Project

Solution: Single Copy Services Many -- Use Data in Place



NOAA Data Services

Data Broker (CICS-NC)

Federal Security Boundary

BDP Collaborators

amazon web services

IBM

Microsoft

Users

Research/ Academia

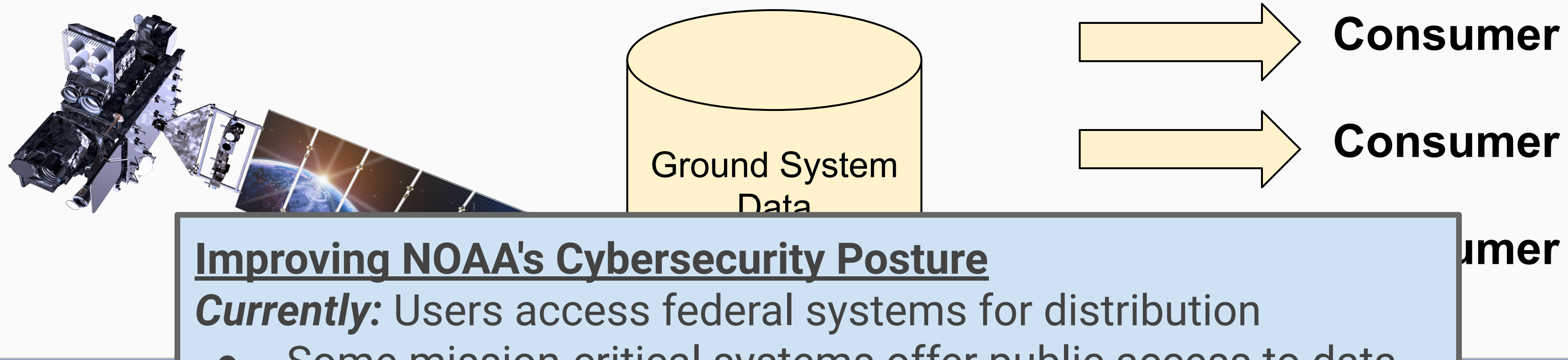
Business

Public

Value-Add Providers

Traditional NOAA Satellite Data Internet Access Strategy

Challenge: Consumer Must Download Data to Use



Improving NOAA's Cybersecurity Posture

Currently: Users access federal systems for distribution

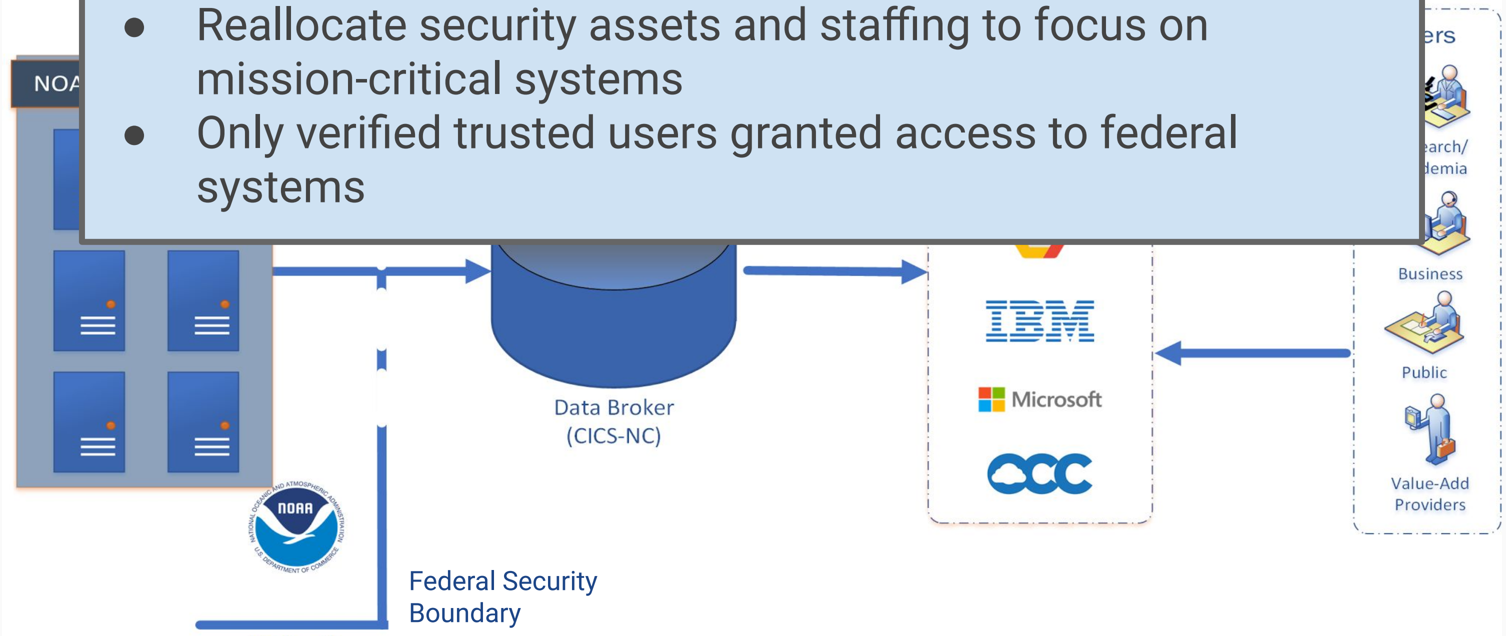
- Some mission-critical systems offer public access to data
- External users have access inside federal security boundary

Potential: All data access occurs in non-federal system

- Reallocate security assets and staffing to focus on mission-critical systems
- Only verified trusted users granted access to federal systems

NOAA

Services
n Place



NOAA data is leveraged through public-private partnerships

CRADA

amazon web services

OCC

Microsoft

Google Cloud Platform

IBM

→

Contract

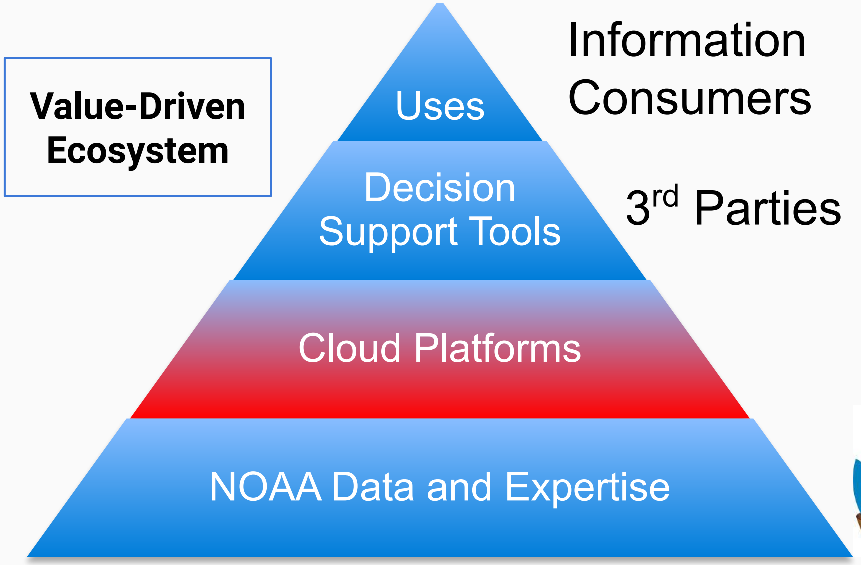
amazon web services

Microsoft

Google Cloud Platform

NOAA:

- Data!
- Science and subject matter expertise
- Cloud data management expertise



NOAA 50 YEARS 1970-2020

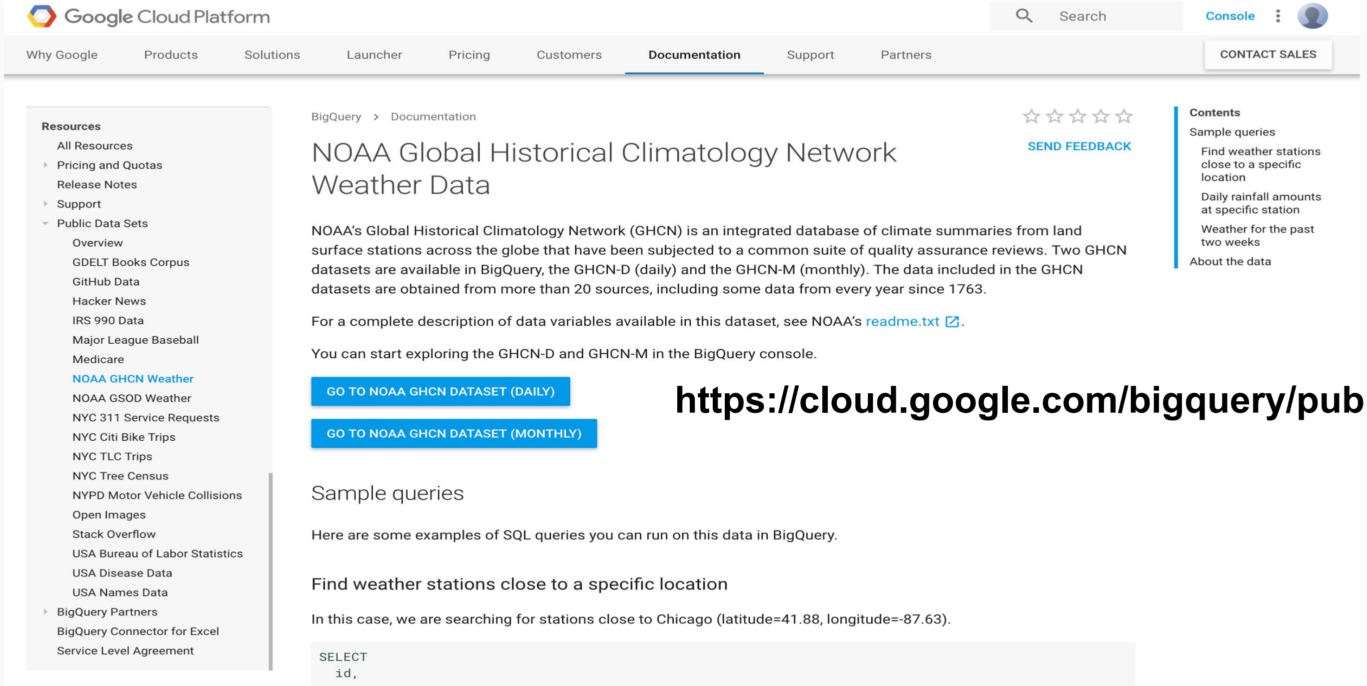
External Partners:

- Provide **FREE** access to the public
- Data storage and access expertise
- Cloud's scalable and on-demand processing capability
- Certify that the data remain open, are not to be sold
- Collaborators monetize services based on data

Dataset Examples

Big Data Project Lessons Learned

Partners’ cloud platforms provide advantages for providing public access to NOAA data.



NOAA Global Historical Climatology Network Weather Data

NOAA’s Global Historical Climatology Network (GHCN) is an integrated database of climate summaries from land surface stations across the globe that have been subjected to a common suite of quality assurance reviews. Two GHCN datasets are available in BigQuery, the GHCN-D (daily) and the GHCN-M (monthly). The data included in the GHCN datasets are obtained from more than 20 sources, including some data from every year since 1763.

For a complete description of data variables available in this dataset, see NOAA’s readme.txt.

You can start exploring the GHCN-D and GHCN-M in the BigQuery console.

Sample queries

Here are some examples of SQL queries you can run on this data in BigQuery.

Find weather stations close to a specific location

In this case, we are searching for stations close to Chicago (latitude=41.88, longitude=-87.63).

```
SELECT
  id,
```

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

Big Data Project Lessons Learned

There is measurable, latent demand for NOAA data.

AWS and Weather Radar

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

Base Reflectivity
NWS Sacramento, CA
National Weather Service WSR-88D Image from: DAX 03/01/2018 22:59 UTC (2:59 PM PST)
Weaverville
Redding
Susanville
Redway
Red Bluff
Chico
Portola
Oroville
Reno
Fort Bragg
Truckee
Ukiah
Marysville
Carson City
Point Arena
Auburn
S Lake Tahoe
Yerington
Woodland
Sacramento
Santa Rosa
Napa
Fairfield
San Rafael
Stockton
Sonora
San Francisco
Modesto
Mammoth
Fremont
San Jose
Merced
Santa Cruz
Hollister
Madera
Salinas
Fresno
Monterey
Hanford
Vcp: 12
DBZ
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
-5
-10
-15
-20
-25
-30
ND
Warnings:
Tornado
Severe Thunderstorm
Flash Flood
Special Marine

☑ Topo ☑ Radar ☑ Counties □ Rivers ☑ Highways ☑ Cities ☑ Warnings ☑ Legend

Following AWS service release:

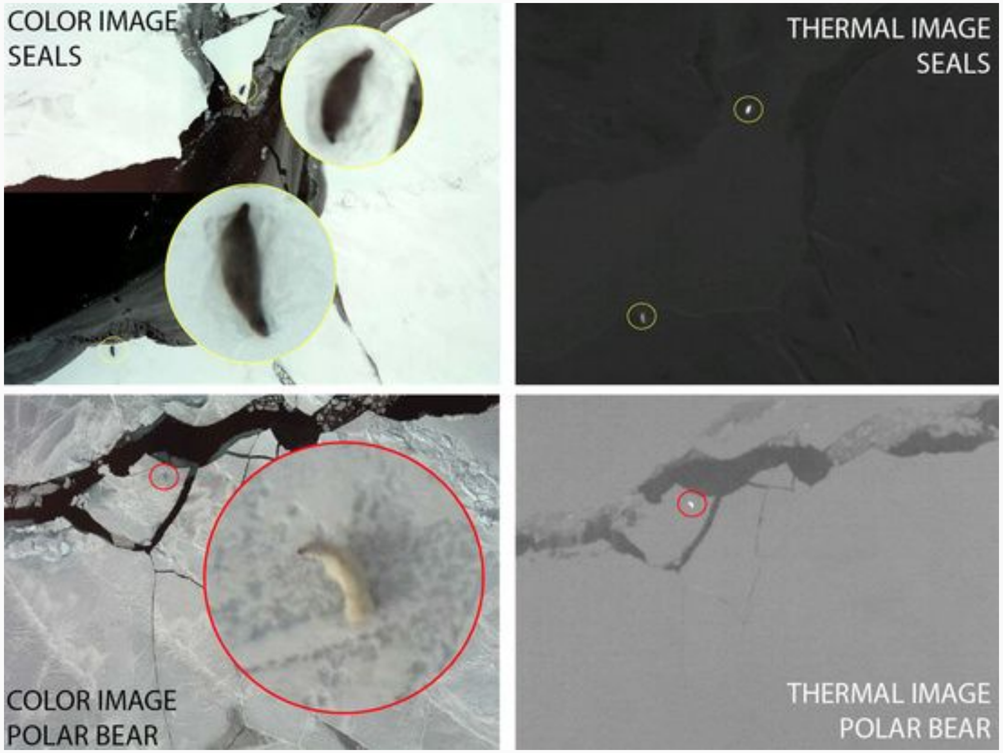
- Increased usage (2.3 times), 50% reduction on NOAA servers.
- New uses – bird migration, mayfly studies
- 80% of NOAA NEXRAD data orders are now served by AWS.
  - *(Ansari et al, 2017 BAMS)*

Big Data Project Lessons Learned

Microsoft AI for Earth

A collaboration between Microsoft AI for Earth, NOAA Fisheries, and the University of Washington, is aimed at automating the detection of arctic seals in aerial imagery. Data are hosted on http://lila.science/datasets/arcticseals, a Microsoft sponsored Azure repository for biology and conservation. Two other preliminary project with Microsoft AI for Earth are:

a. Automating the detection of beluga whale sounds in hydrophone data.
b. Automating the detection and classification of fish in video.



Further information:
Benefits & how it works

Big Data Project
Lessons Learned and Best Practices

- **The concept is technically feasible & improves NOAA's security posture**
- **Partners' cloud platforms have significant advantages for providing access**
- **Integrating NOAA data into tools has the greatest usage potential**
- **The scarce asset that NOAA brings to the BDP is its expertise**
- **NOAA owns data quality, and should establish a means of authentication to establish public trust in cloud-based data**
- **Cloud-optimized formats represent a challenge and a potential advantage for cloud data process and analytics. At present, there is not broad agreement as to which format(s) (Zarr, Parquet, Cloud-optimized GeoTIFF, Xarray, etc.) are the most in demand**

Collaborative Solutions

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

BEING A CLOUD WAS SUCH A CAREFREE EXISTENCE – NOW PEOPLE EXPECT SO MUCH FROM US!
© D.Fletcher for CloudTweaks.com

Big Data Project Lessons Learned

- **The role of an intermediate "data broker" has emerged as a valuable function & possible Enterprise Service that could support NOAA in provision of data to the commercial cloud.**
    - **The broker can extract a single copy of data from NOAA's existing access services and systems and send the data to multiple Collaborators in a reliable and quality-assured way, thus saving NOAA's bandwidth and effort.**
    - **NOAA may effectively improve its IT security posture through the publishing of its data to the commercial cloud for general public access. Only a trusted agent needs to access NOAA federal data services and re-distribute to a non-federal system for access, thus greatly reducing risk and exposure on the federal system**

Further information:
History of the contract

BDP CRADA Basics

amazon web services

OCC

Microsoft

Google Cloud Platform

IBM

- April 2015: NOAA announced five cooperative research and development agreements (CRADAs) between NOAA and selected commercial and academic collaborators
    - 5 separate but identical 4-year agreements

- Selected datasets of interest to the BDP collaborators and their business partners are being transferred from NOAA to the collaborators' infrastructure.

- The partnership has already realized a number of dataset improvements, and new and novel applications that employ NOAA's atmospheric, geophysical, and oceanographic data, at no net cost to the US taxpayer.

Big Data Project RFI Results

On October 1st 2018, the National Oceanic and Atmospheric Administration (NOAA) published in the Federal Register a Request for Information (RFI) with an objective of informing the future direction of the Big Data Project. 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.

- Responses were positive about the access users had to NOAA data
- Users wanted more assurance that data would be available over a longer period of time, rather than just during an experimental phase.
- Cloud Service Providers expressed a willingness to continue under a longer term agreement

The New BDP Vehicle: RFP Description

- **2 year base IDIQ Contract, 4x2 option periods**
  - **New Feature** in response to Industry and NOAA comments regarding commitment
- **Rules of engagement are the same as CRADA phase**
  - Open data/no charge for egress/charge for services and products built on data
- **NOAA Allocation request of NLT 5 PB of storage**
  - Data included to be determined by NOAA
  - **New feature** in response to NOAA needs
- **Partners store data and provide access for free**
  - From the NOAA 5 PB Allocation, and/or
  - Partners may also request any NOAA open data, and/or
  - Any NOAA data that are being stored by NOAA on the Partners' commercial cloud platforms using other contract vehicles for mission purposes (free egress)

Big Data Program Contract Awardees

- **AWS**
  - https://aws.amazon.com/noaa-big-data/
- **Google Cloud Platform**
  - https://cloud.google.com/bigquery/public-data/
  - https://explorer.earthengine.google.com/#index
- **Microsoft**
  - https://azure.microsoft.com/en-us/services/open-datasets/catalog/

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/TIME WELLSPENT