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/](https://aws.amazon.com/noaa-big-data/)

- **Google Cloud Platform**
  - [https://cloud.google.com/bigquery/public-data/](https://cloud.google.com/bigquery/public-data/)
  - [https://explorer.earthengine.google.com/#index](https://explorer.earthengine.google.com/#index)

- **IBM**
  - [https://noaa-crada.mybluemix.net/](https://noaa-crada.mybluemix.net/)

- **Microsoft**

- **Open Commons Consortium**
Why
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

Ground System Data Distribution

Consumer
Consumer
Consumer
Consumer
Consumer
One-way transfer out of federal systems. Only a **trusted user** inside security boundary.

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

Using it is the hardest part.
Value Driven Ecosystem

- NOAA Data and Expertise
- Cloud Platforms
- Decision Support Tools
- Uses

Information Consumers
3rd Parties

[Logos of cloud providers and other companies]
● Weekly/bi-weekly/monthly, per collaborator
● NOAA “data pitches” to collaborators
● Communications/Scheduling/Documenting
● Technical and non-technical NOAA liaison staff
● BDP Advisory Board
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

● Open Commons Consortium
  ○ http://edc.occ-data.org/
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.
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
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

- **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

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 1793.

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.

https://cloud.google.com/bigquery/public-data/noaa-ghcn
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/hwmuE5UwK

10:11 AM - Sep 1, 2017
Data Broker’s Usage Statistics Analysis (presented at Earth Science Information Partners (ESIP) meeting)
BDP Data Broker Update

Jonathan Brannock and Otis Brown
CISESS / NCICS
NCSU
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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Cloud Provider(s)</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOES-16/17 ABI &amp; GLM</td>
<td>~1 TB daily</td>
<td>AWS, GCS, OCC</td>
<td>NESDIS/ESPC</td>
</tr>
<tr>
<td>GOES-15 GVAR</td>
<td>~15 GB daily</td>
<td>OCC</td>
<td>NESDIS/ESPC</td>
</tr>
<tr>
<td>NEXRAD L2</td>
<td>~60 - 201 GB daily</td>
<td>AWS, GCS, OCC</td>
<td>NESDIS/NCEI</td>
</tr>
<tr>
<td>NEXRAD L3</td>
<td>~20 - 60 GB daily</td>
<td>GCS</td>
<td>NESDIS/NCEI</td>
</tr>
<tr>
<td>GHCN-D</td>
<td>~30 GB daily</td>
<td>AWS</td>
<td>NESDIS/NCEI</td>
</tr>
<tr>
<td>National Water Model</td>
<td>~115 GB daily</td>
<td>AWS, GCS</td>
<td>NWS/NCO</td>
</tr>
<tr>
<td>Ocean Forecast System</td>
<td>~288 GB daily</td>
<td>AWS</td>
<td>NWS/NCO</td>
</tr>
</tbody>
</table>
## Brokered Datasets (Continued)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Cloud Provider(s)</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOAA-20 VIIRS DNB</td>
<td>~60 GB daily</td>
<td>AWS</td>
<td>NESDIS/ESPC</td>
</tr>
<tr>
<td>Fire Products</td>
<td>~1.1 GB daily</td>
<td>AWS,GCS</td>
<td>NESDIS/ESPC</td>
</tr>
<tr>
<td>ISD</td>
<td>~750 MB daily</td>
<td>AWS</td>
<td>NESDIS/NCEI</td>
</tr>
<tr>
<td>Global Hourly</td>
<td>~20 GB daily</td>
<td>AWS</td>
<td>NESDIS/NCEI</td>
</tr>
<tr>
<td>GFS</td>
<td>~950 GB daily</td>
<td>AWS</td>
<td>NOS/CO-OPS</td>
</tr>
<tr>
<td>HRRR</td>
<td>~920 GB daily</td>
<td>AWS,GCS</td>
<td>NWS/NCO</td>
</tr>
<tr>
<td>CFS</td>
<td>~310 GB daily</td>
<td>AWS</td>
<td>NWS/NCO</td>
</tr>
<tr>
<td>Dataset</td>
<td>Count</td>
<td>Dataset</td>
<td>Bytes</td>
</tr>
<tr>
<td>------------------</td>
<td>----------------</td>
<td>-------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>GOES-16</td>
<td>754,659,402</td>
<td>GOES-16</td>
<td>3,163,532GB</td>
</tr>
<tr>
<td>GOES-17</td>
<td>80,221,671</td>
<td>GOES-17</td>
<td>584,946GB</td>
</tr>
<tr>
<td>NWM (Archive)</td>
<td>11,530,076</td>
<td>NWM (Archive)</td>
<td>142,768GB</td>
</tr>
<tr>
<td>GFS</td>
<td>3,048,159</td>
<td>GHCNd</td>
<td>50,658GB</td>
</tr>
<tr>
<td>GHCNd</td>
<td>1,116,821</td>
<td>GFS</td>
<td>41.800GB</td>
</tr>
<tr>
<td>NWM</td>
<td>136,287</td>
<td>GEFS</td>
<td>.377GB</td>
</tr>
<tr>
<td>ISD</td>
<td>114,879</td>
<td>GFS(para)</td>
<td>.297GB</td>
</tr>
<tr>
<td>GEFS</td>
<td>63,598</td>
<td>OFS</td>
<td>.066GB</td>
</tr>
</tbody>
</table>

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]

Accesion Ratio

<table>
<thead>
<tr>
<th>Time (Months)</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/1/18</td>
<td>0</td>
</tr>
<tr>
<td>1/1/19</td>
<td>0</td>
</tr>
<tr>
<td>2/1/19</td>
<td>5</td>
</tr>
<tr>
<td>3/1/19</td>
<td>5</td>
</tr>
<tr>
<td>4/1/19</td>
<td>10</td>
</tr>
<tr>
<td>5/1/19</td>
<td>5</td>
</tr>
<tr>
<td>6/1/19</td>
<td>25</td>
</tr>
</tbody>
</table>
GOES-16/17 Data Statistics [AWS]
GOES-16/17 Data Statistics [AWS]
Age is the time from observation. Average frequencies are: ~50%, ~18%, ~8% and ~24%, respectively.
Data on Cloud Platforms
STI/R2O Use Cases
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 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
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)](https))--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?

SEE, I TOLD YOU THAT BIG DATA WAS TOO SCARY

TIME WELL SPENT™
by Tom Fishburne

© 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