Data on the Web

https://opendata.aws

Jed Sundwall, Global Open Data Lead

The project started with the philosophy that much academic information should be freely available to anyone. It aims to allow information sharing within internationally dispersed teams, and the dissemination of information by support groups.

[...]

The WWW world consists of documents, and links.

— Tim Berners-Lee in 1991

https://www.w3.org/People/Berners-Lee/1991/08/art-6487.txt

Man plans and God laughs.

— Yiddish proverb

Cloud computing is the on-demand delivery IT resources via the internet with pay-as-you-go pricing.

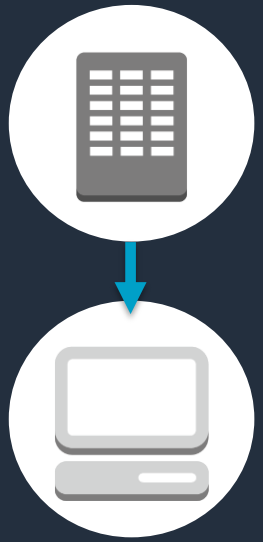
Sharing data in the cloud lets data users spend more time on data analysis rather than data acquisition.

https://opendata.aws

Flipped data flow in the cloud

Traditional approach:
Move the data to computing resources.

Cloud approach:
Move computing resources to the data.



Advantages of sharing data in the cloud

Global community of users

New services and tools

Faster pace of research

Lower cost of research

Take full advantage of the web!

AWS Public Datasets

https://registry.opendata.aws

AWS Public Datasets

https://registry.opendata.aws/tag/satellite-imagery

registry.opendata.aws/tag/satellite-imagery/

Registry of Open Data on AWS

About

This registry exists to help people discover and share datasets that are available via AWS resources. Learn more about sharing data on AWS.

See all usage examples for datasets listed in this registry tagged with **satellite imagery**.

Search datasets (currently 10 matching datasets)

Search datasets

You are currently viewing a subset of data tagged with **satellite imagery**.

Add to this registry

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the Registry of Open Data on AWS GitHub repository.

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third parties under a variety of licenses. Please check dataset licenses and related documentation to determine if a dataset may be used for your application.

Sentinel-2

earth observation | satellite imagery | gis | natural resource | sustainability | disaster response

The Sentinel-2 mission is a land monitoring constellation of two satellites that provide high resolution optical imagery and provide continuity for the current SPOT and Landsat missions. The mission provides a global coverage of the Earth's land surface every 5 days, making the data of great use in on-going studies. L1C data are available from June 2015 globally. L2A data are available from April 2017 over wider Europe region and globally since December 2018.

Details →

Usage examples

- Sentinel Hub WMS/WMTS/WCS Service by Sinergise
- Python package for working with Sentinel-2 AWS data by Sinergise
- Exploring the Chile wildfires with Landsat and Sentinel-2 imagery by Timothy Whitehead
- Sentinel Playground by Sinergise
- EOS Land Viewer by Earth Observing System

See 16 usage examples →

Landsat 8

earth observation | satellite imagery | gis | natural resource | sustainability | disaster response

An ongoing collection of satellite imagery of all land on Earth produced by the Landsat 8 satellite.

Details →

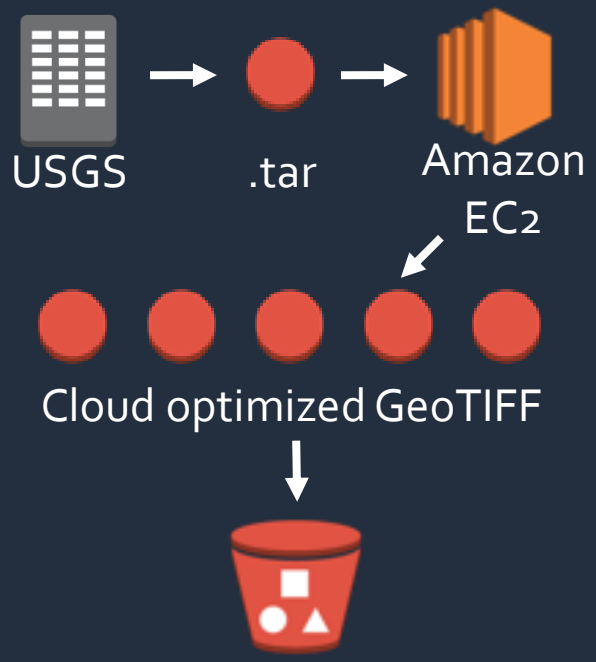
Usage examples

- Sentinel Playground for Landsat by Sinergise
- Integrate imagery from the full Landsat archive into your own apps, maps, and

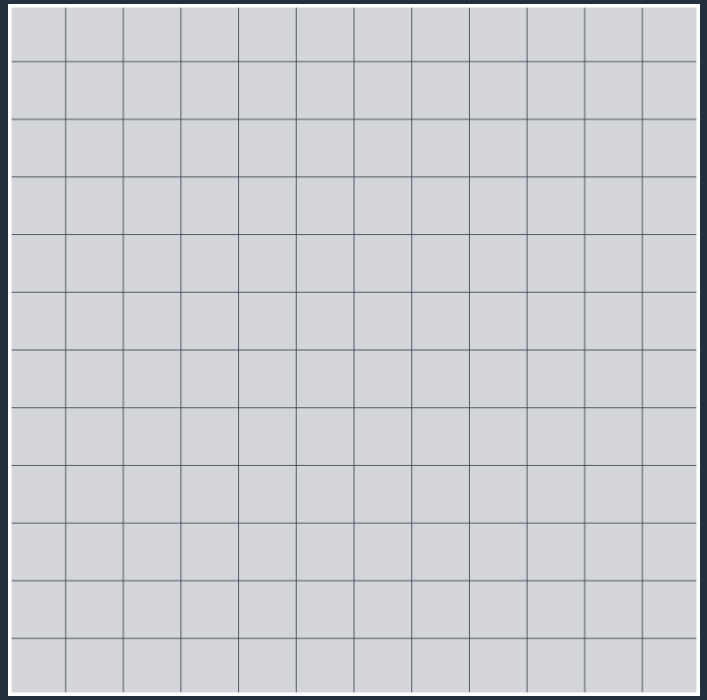
Earth on AWS

aws.amazon.com/earth

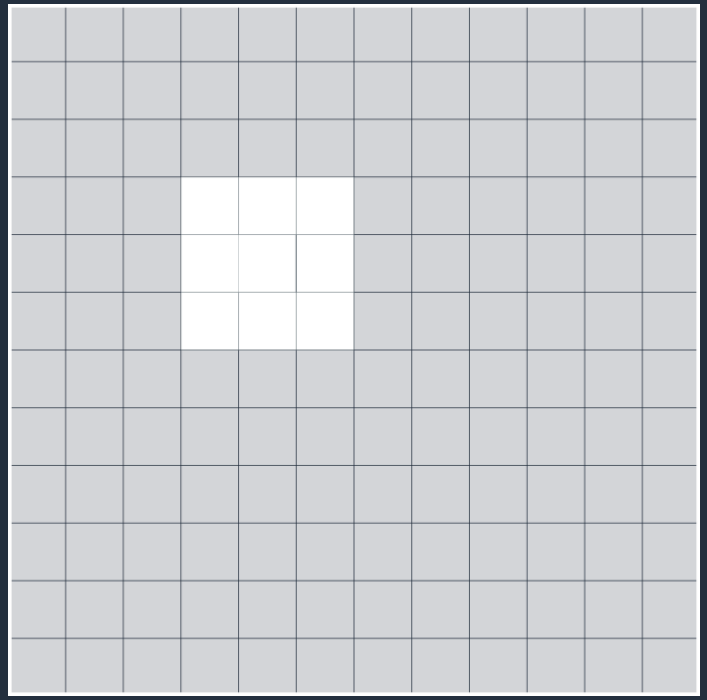
Staging data for analysis



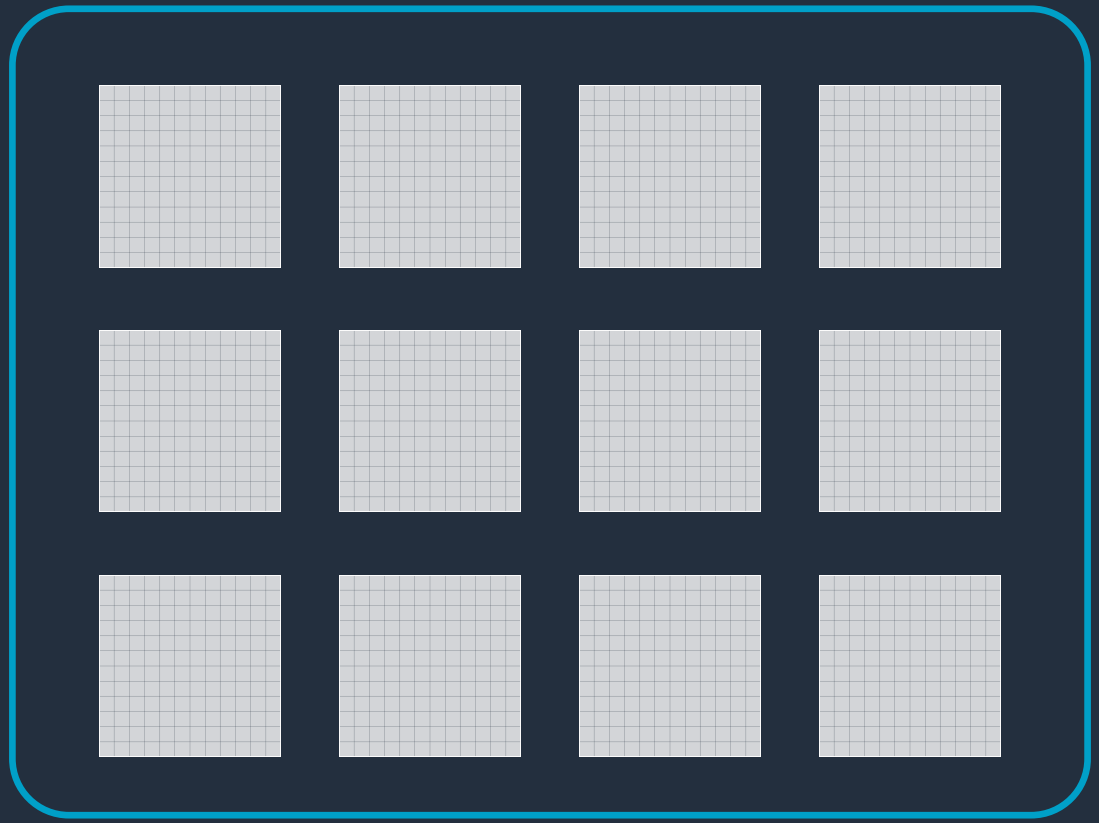
The cloud-optimized GeoTIFF



The cloud-optimized GeoTIFF

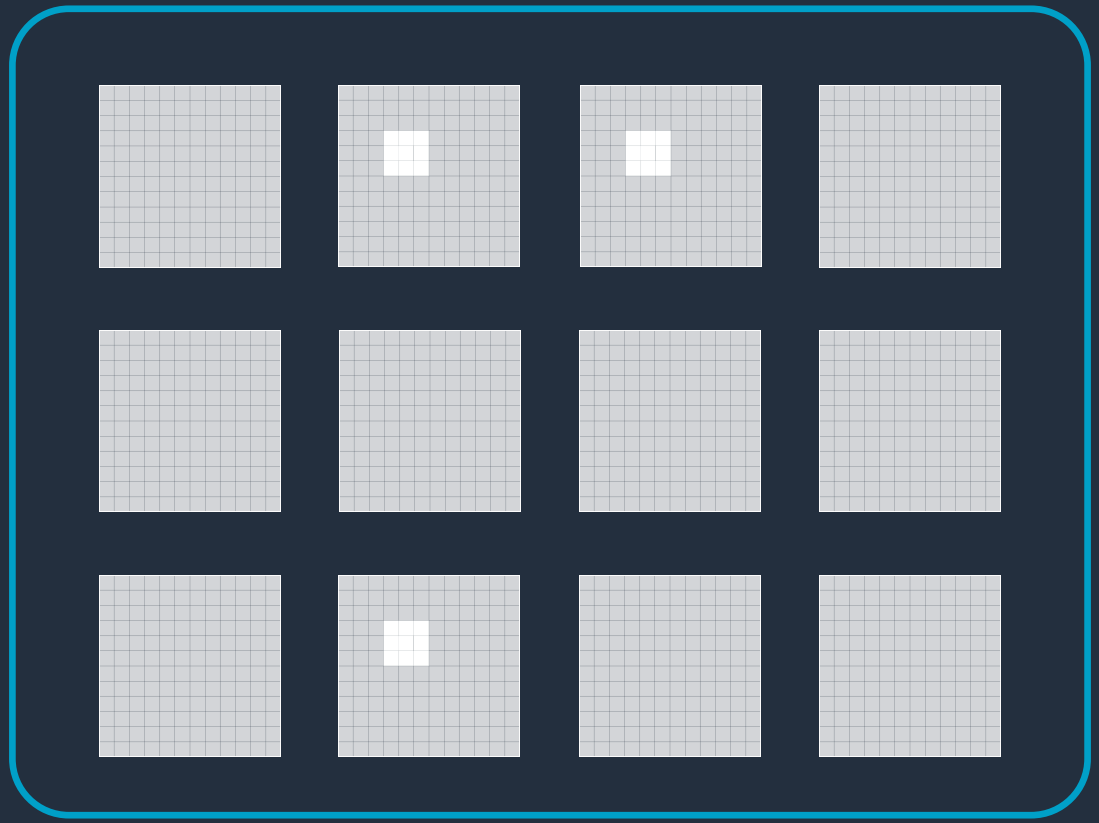


The cloud-optimized GeoTIFF



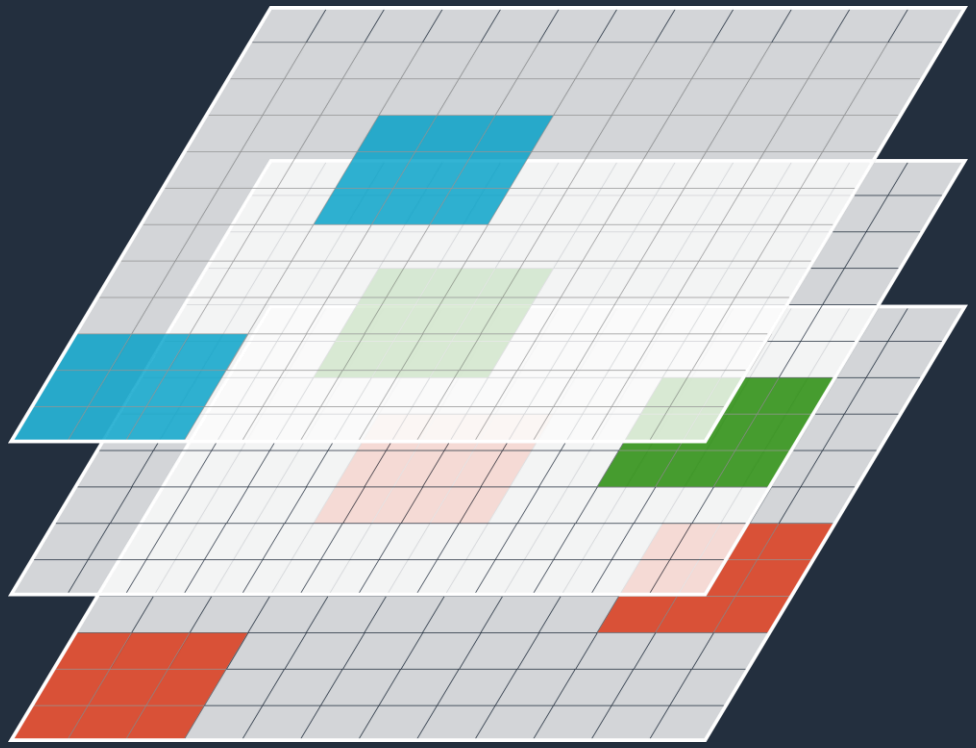
.tar

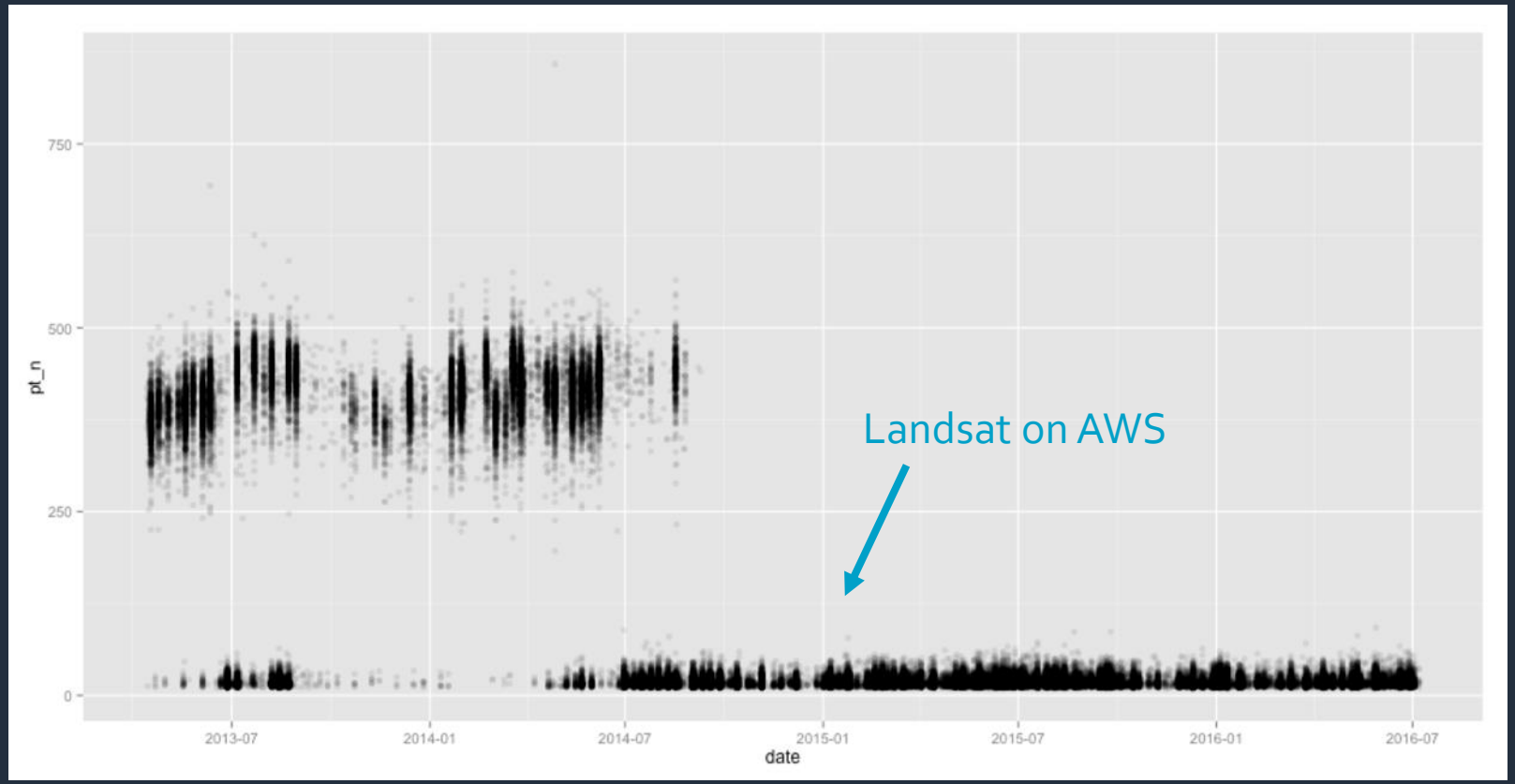
The cloud-optimized GeoTIFF



.tar

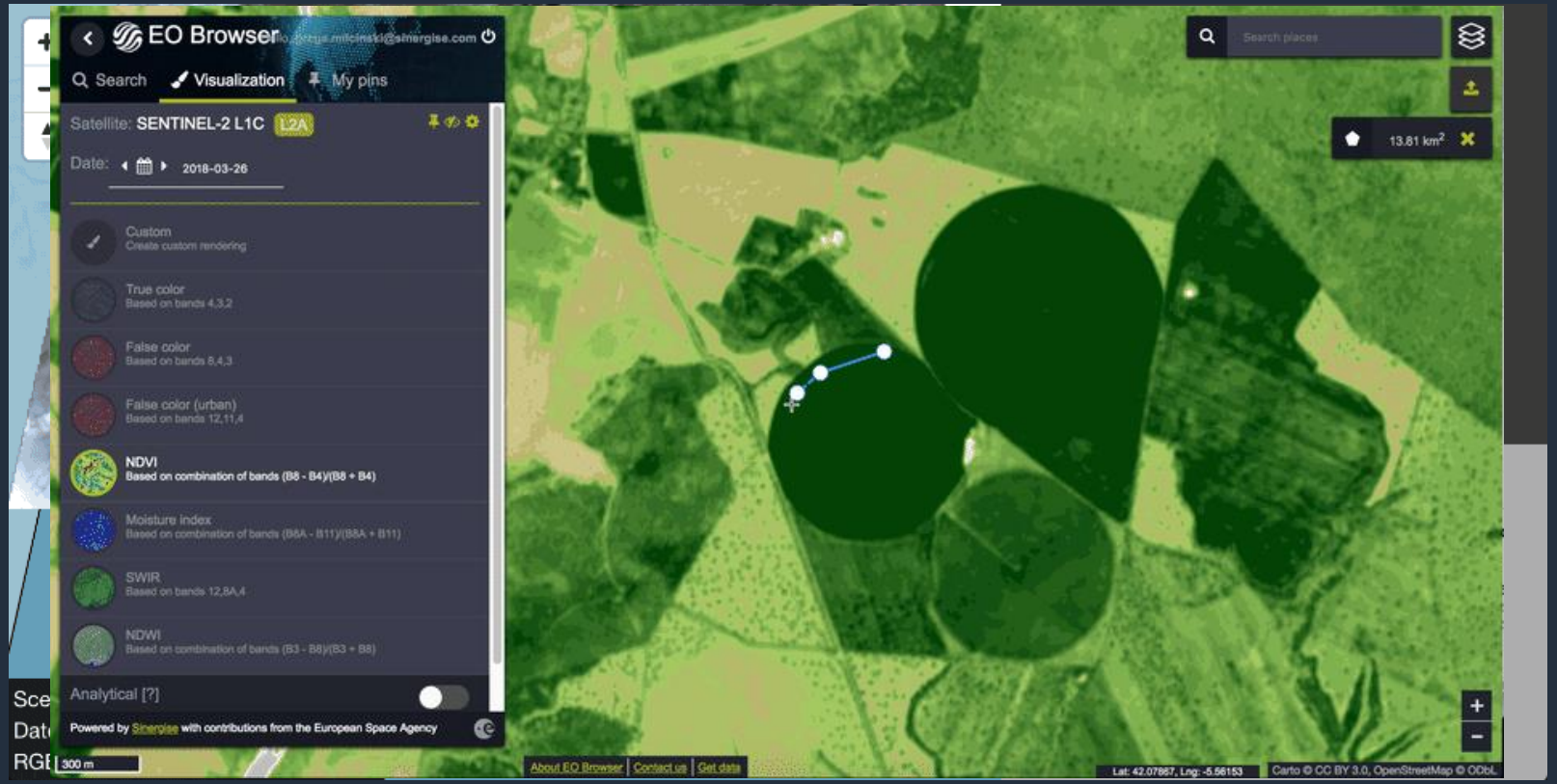
The cloud-optimized GeoTIFF





Graph by Drew Bollinger (@drewb019) at Development Seed

Using Serverless to Visualize and Analyze Imagery



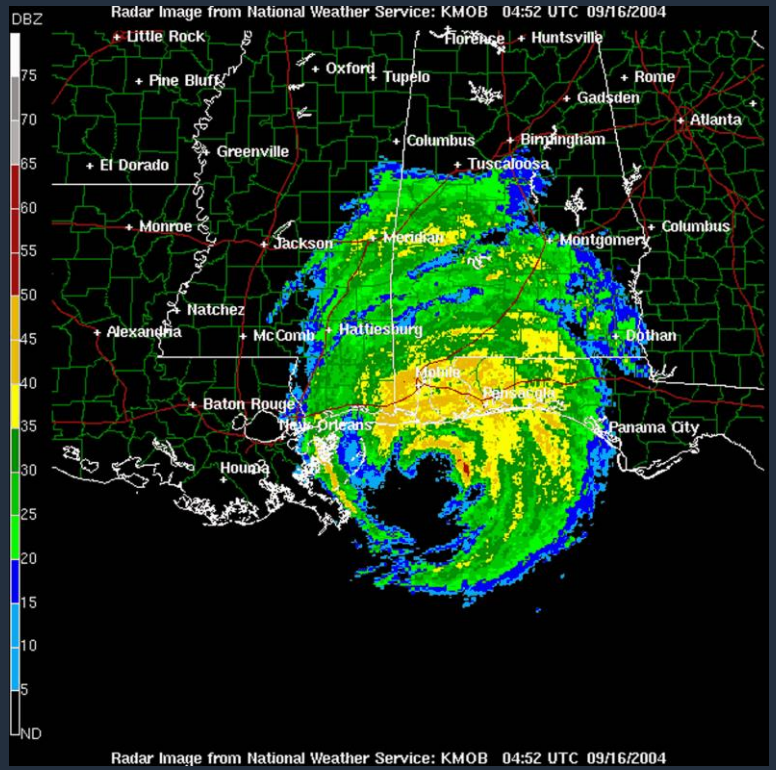
NEXRAD on AWS

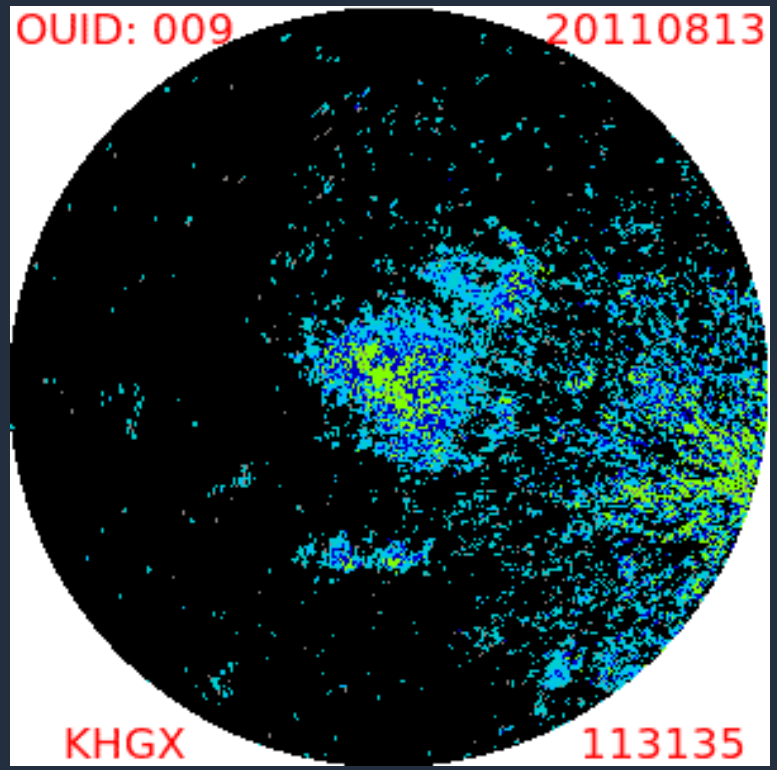
"Recently, the National Oceanic and Atmospheric Administration and Amazon Web Services (AWS) Cloud made available one of the largest datasets describing animal movement ever compiled: …"

— Adriaan M. Dokter et al. Nature (2018)

"Recently, the National Oceanic and Atmospheric Administration and Amazon Web Services (AWS) Cloud made available one of the largest datasets describing animal movement ever compiled: the Next Generation Weather Radar (NEXRAD) archive."

— Adriaan M. Dokter et al. Nature (2018)

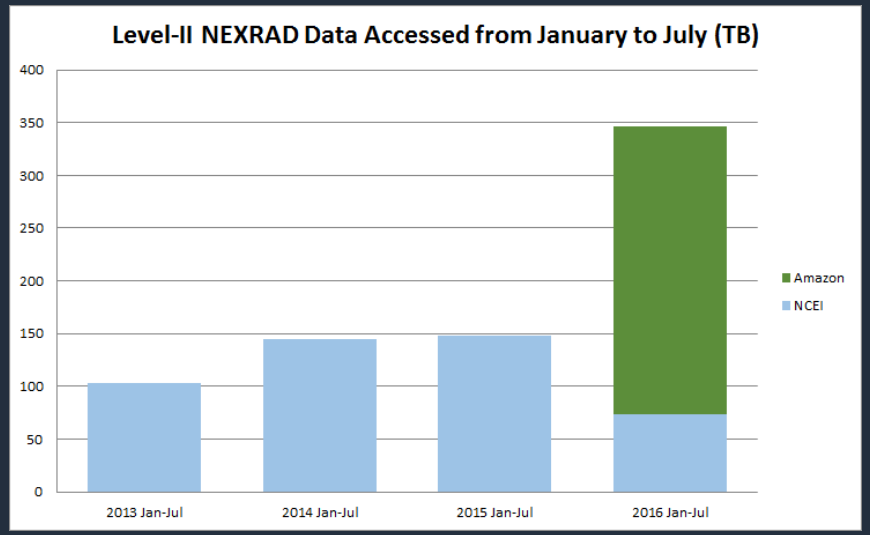




NEXRAD on AWS

Immediate usage:

- Climate Corporation **cut two weeks** out of an analysis pipeline
- Increased NEXRAD usage **2.3x**
- A weather data company stopped storing their own NEXRAD archive, **freeing up revenue** to build new products.



openaq

"Our mission is to enable previously impossible science, impact policy and empower the public to fight air pollution through open data, open-source tools, and cooperation."

https://openaq.org

aws

What makes a dataset successful?

It is treated like a product.

Common Crawl - Registry of O

Guest

Secure | https://registry.opendata.aws/commoncrawl/

Registry of Open Data on AWS

aws

Common Crawl

encyclopedic | machine learning | internet

Description

A corpus of web crawl data composed of over 5 billion web pages.

Update Frequency

Monthly

License

This data is available for anyone to use under the Common Crawl Terms of Use

Documentation

http://commoncrawl.org/the-data/get-started/

Contact

http://commoncrawl.org/connect/contact-us/

Usage Examples

- Building a Web-Scale Dependency-Parsed Corpus from CommonCrawl by Alexander Panchenko, et al.
- Dresden Web Table Corpus (DWTC) by Database Systems Group Dresden
- Index to WARC Files and URLs in Columnar Format by Sebastian Nagel

Resources on AWS

Description
Crawl data (WARC and ARC format)

Resource type
S3 Bucket

Amazon Resource Name (ARN)
`arn:aws:s3:::commoncrawl`

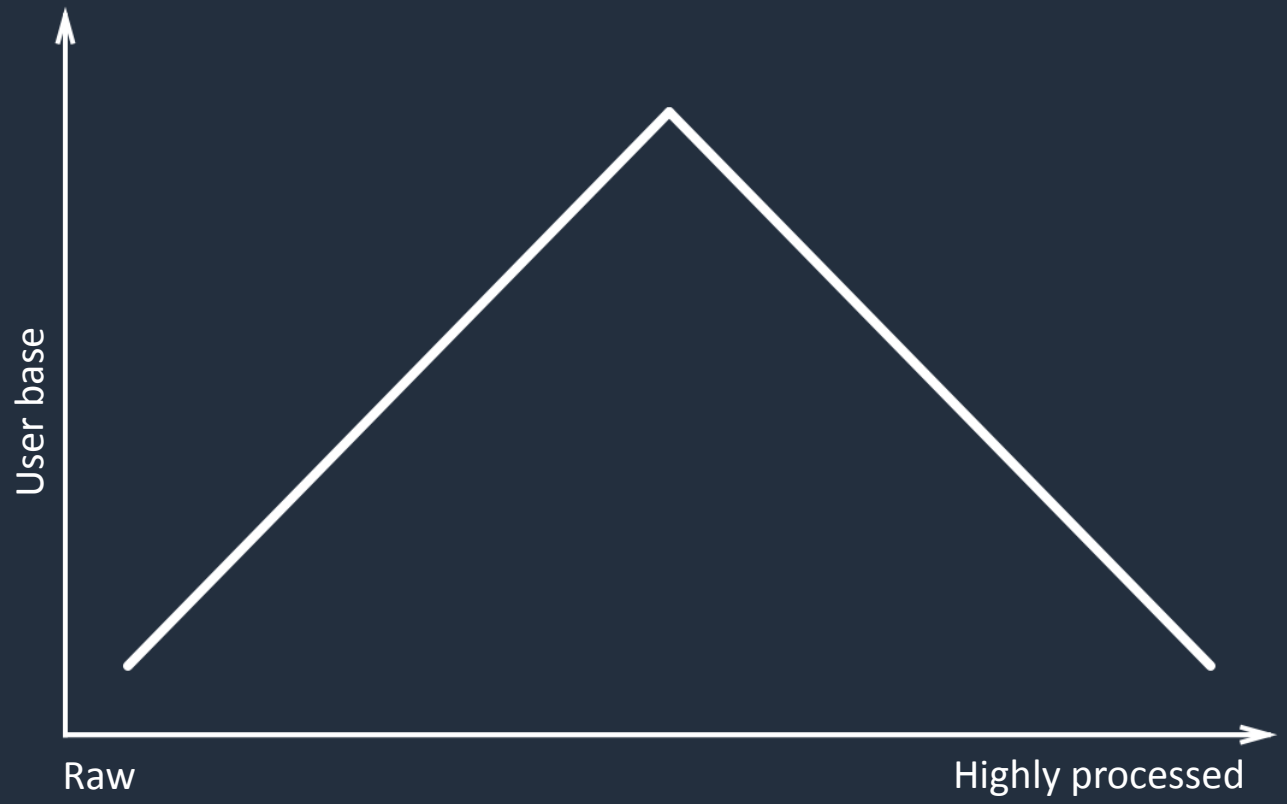
AWS Region
`us-east-1`

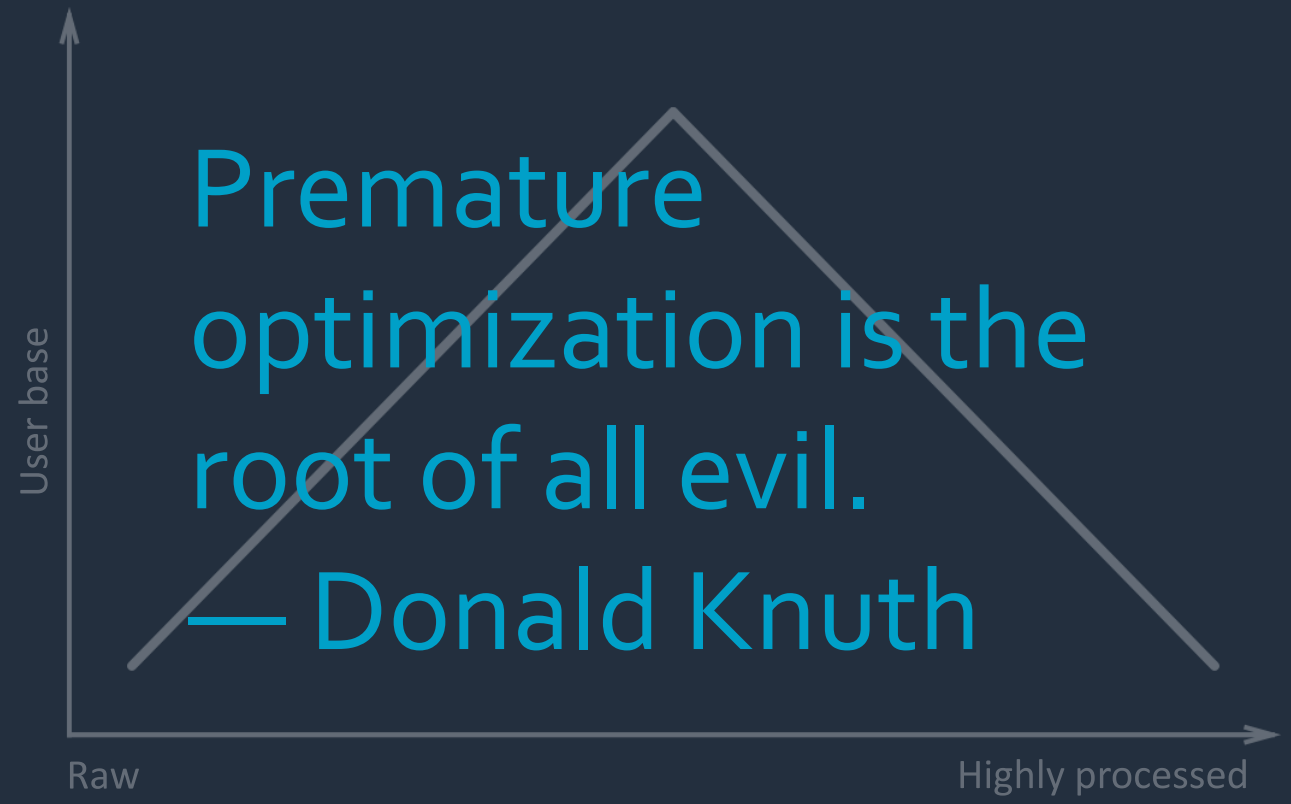
aws

What makes a dataset successful?

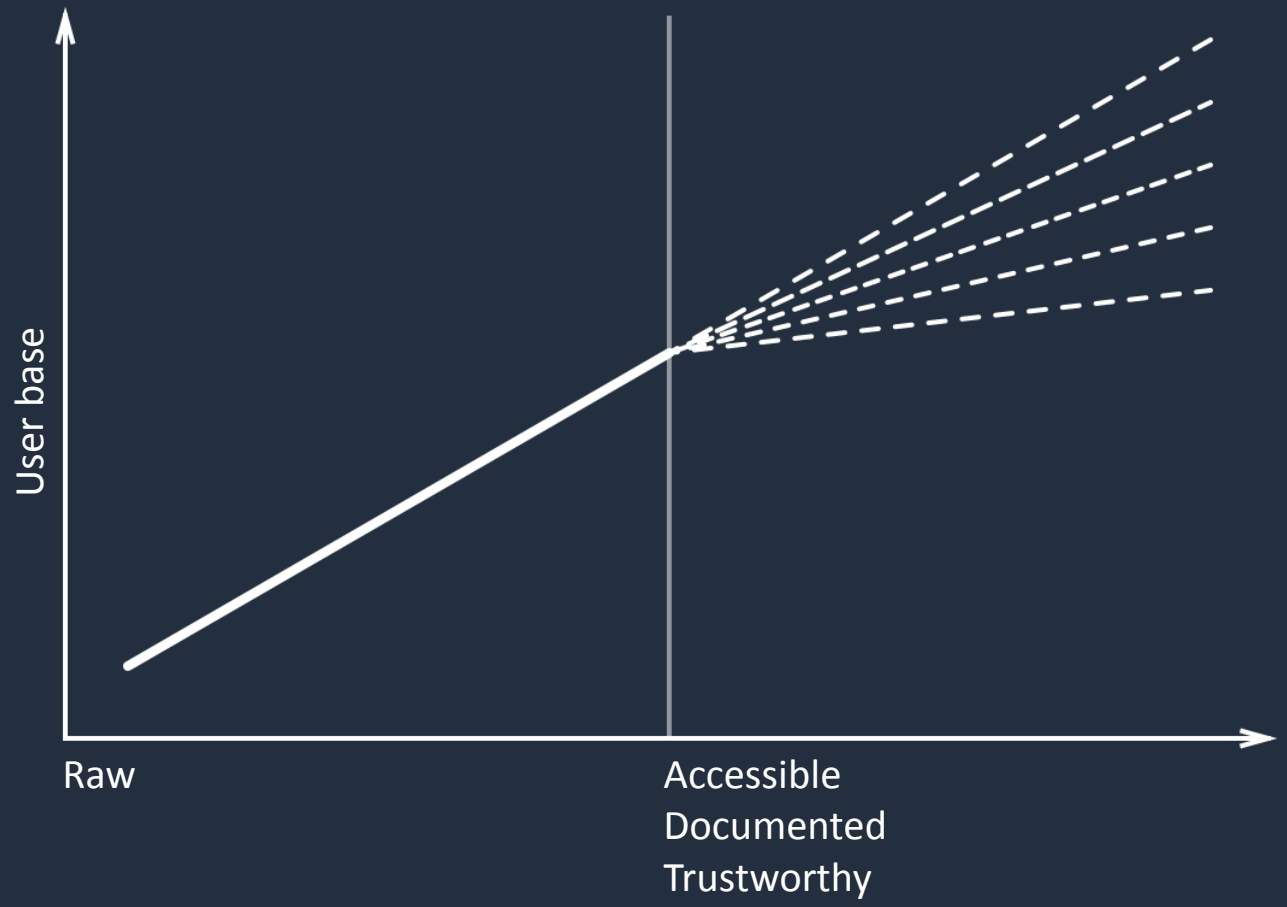
It is treated like a product.

It is optimized for analysis.

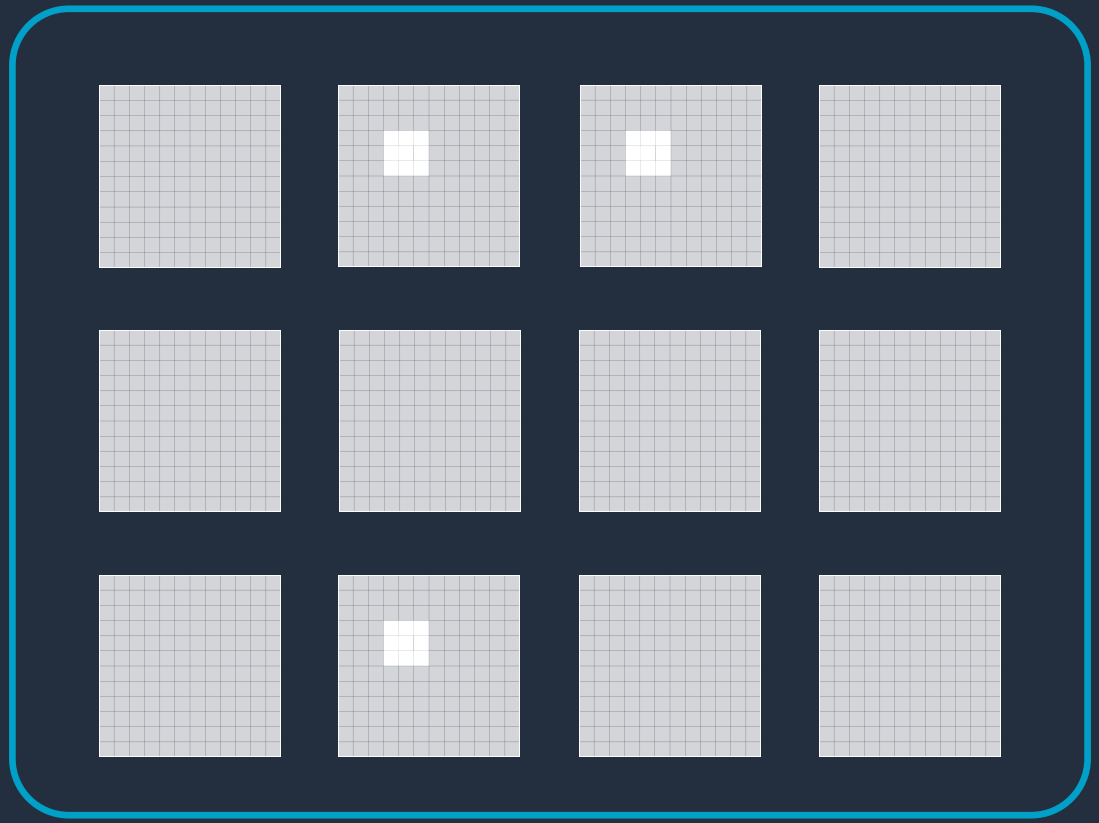




Premature optimization is the root of all evil.
— Donald Knuth

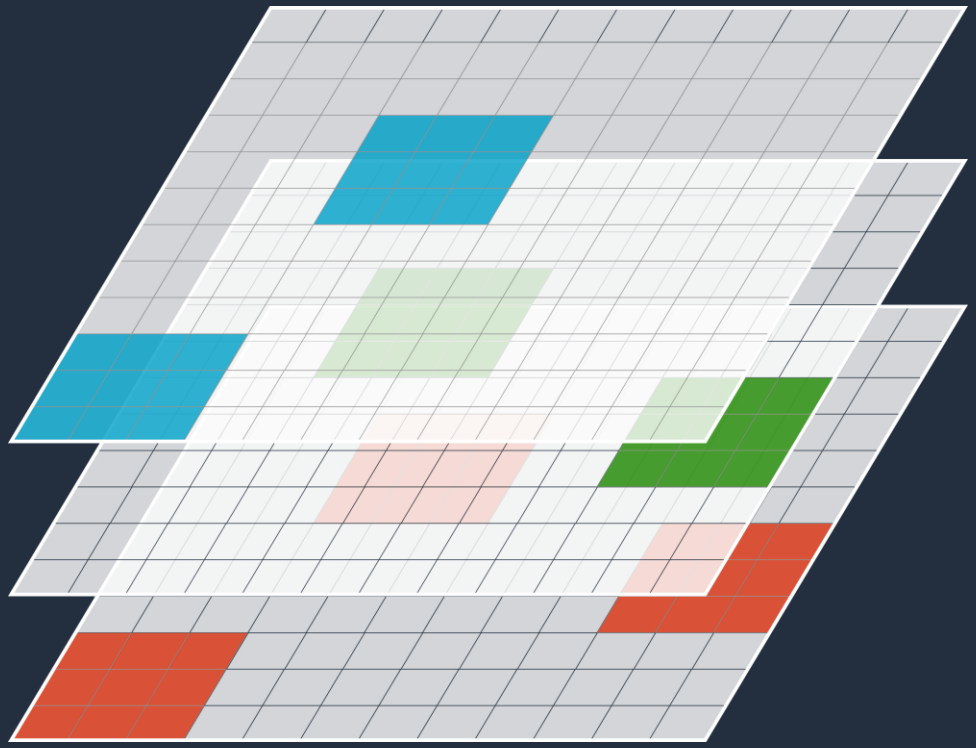


The cloud-optimized GeoTIFF



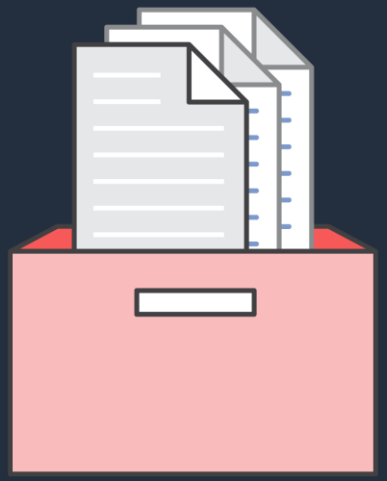
.tar

The cloud-optimized GeoTIFF



Patterns

S3 Key Index



External Index

Internal Index

Example: GOES-16 Key Naming

```
s3://noaa-goes16/ABI-L1b-RadF/2018/149/14/
OR_
ABI-L1b-RadF-M3C14_
G16_
s20181491430465_
e20181491441232_
c20181491441300.nc
```

Example: IRS 990 CSV as External Index

Book1

Home Insert Page Layout Formulas Data Review View Share

A2 15109264

	A	B	C	D	E	F	G	H	I
1	RETURN_ID	FILING_TYPE	EIN	TAX_PERIOD	SUB_DATE	TAXPAYER_NAME	RETURN_TYPE	DLN	OBJECT_ID
2	15109264	EFILE	453578215	201612	1/10/18 13:03	MULEY FANATIC FOUNDATION OF WY	990	93493318071517	201713189349307000
3	15109263	EFILE	383333202	201612	1/10/18 13:03	KALAMAZOO COMMUNITY FOUNDATI	990	93493318071467	201713189349307000
4	15109260	EFILE	233014323	201612	1/10/18 13:03	GOSPEL THROUGH COLOMBIA	990	93493318071317	201713189349307000
5	15109257	EFILE	351837569	201612	1/10/18 13:03	PREMIER ARTS INC	990	93493318071117	201713189349307000
6	15109256	EFILE	133135292	201706	1/10/18 13:03	ELDERS SHARE THE ARTS INC	990	93493318071067	201713189349307000
7	15109253	EFILE	463224351	201612	1/10/18 13:03	US MILITARY SUPPORT GROUP INC	990	93493318070867	201713189349307000
8	15109246	EFILE	421122161	201706	1/10/18 13:03	PROGRESS INDUSTRIES	990	93493318043117	201713189349304000
9	15109245	EFILE	160983042	201612	1/10/18 13:03	EAST HILL FAMILY MEDICAL INC	990	93493318043067	201713189349304000
10	15109302	EFILE	721483958	201612	1/10/18 13:12	PARKING FACILITIES CORPORATION	990	93493317081567	201713179349308000
11	15109300	EFILE	770201505	201612	1/10/18 13:12	SANTA BARBARA WILDLIFE CARE NETV	990	93493317081467	201713179349308000
12	15109299	EFILE	237439392	201612	1/10/18 13:12	IDAHO LAW FOUNDATION INC	990	93493317081367	201713179349308000
13	15109297	EFILE	860654061	201612	1/10/18 13:12	SIERRA MADRE ALLIANCE INC	990	93493317081167	201713179349308000
14	15108190	EFILE	416027765	201612	1/10/18 10:17	GREYSTONE FOUNDATION	990PF	93491320003067	201713209349100000
15	15108187	EFILE	464902444	201512	1/10/18 10:17	ALAN AND GAIL COHN FOUNDATION II	990PF	93491319023057	201703199349102000
16	15108185	EFILE	271658370	201612	1/10/18 10:17	PHINNEY CHARITABLE FOUNDATION C	990PF	93491319022957	201703199349102000
17	15108181	EFILE	943400451	201612	1/10/18 10:17	OLANDER FAMILY FOUNDATION INC	990PF	93491319022757	201703199349102000
18	15108177	EFILE	463971698	201612	1/10/18 10:17	HJ-99 FOUNDATION	990PF	93491319022557	201703199349102000
19	15108204	EFILE	464381406	201511	1/10/18 10:17	LOYINKJ FOUNDATION	990PF	93491320006017	201713209349100000

Sheet1

Ready 100%

What makes a dataset successful?

It is treated like a product.

It is optimized for analysis.

There is a community around it.

ESIP Summer 2018 Meeting

Earth Science Information Partners

azavea

PANGEO

CORNELL UNIVERSITY FOUNDED A.D. 1865

The HDF Group

PLANET OS

Element 84

Cooperative Institute for Climate and Satellites cics.nc

esri®

aws

Many users = many use cases and tools

ESIP Summer 2018 Meeting

The universe of meteorological data users is expanding to include:

- Economists
- Software developers (web and app developers)
- Young students
- Amateurs

These users have different:

- Skills
- Tools
- Needs

Big is different

ESIP Summer 2018 Meeting

- Object storage is different than file storage
- Toolmakers must keep up with emerging formats
- A number of cloud-friendly formats are emerging
    - Cloud-optimized GeoTIFF (cogeo.org)
    - Zarr
    - NetCDF to Parquet/ORC

Thank you!

jed@amazon.com