Learnings from Staging Petabytes of Data for Analysis in AWS

Joe Flasher, Open Geospatial Data Lead

Cloud computing is the on-demand delivery of compute power, database, storage, applications, and other IT resources via the internet with pay-as-you-go pricing.

Traditional Infrastructure

- Equipment
- Resources and Administration
- Contracts
- Cost

AWS Cloud

- No Up Front Expense Pay for what you Use
- Improve Time to Market & Agility
- Scale Up and Down
- Self-Service Infrastructure

Why does AWS care about open data?

Sharing data on AWS makes it accessible to a large and growing community of researchers, entrepreneurs, and enterprises who use the AWS Cloud.

Many AWS customers supply data to the public to accelerate research and product development.

Many AWS customers use data shared on AWS to create new products and services.

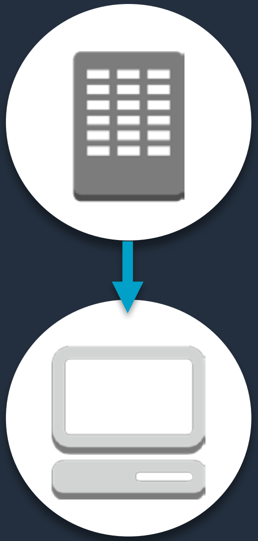
Sharing data in the cloud lets data users spend more time on data analysis rather than data acquisition.

https://opendata.aws

Flipped data flow in the cloud

Traditional approach:

Move the data to computing resources.



Cloud approach:

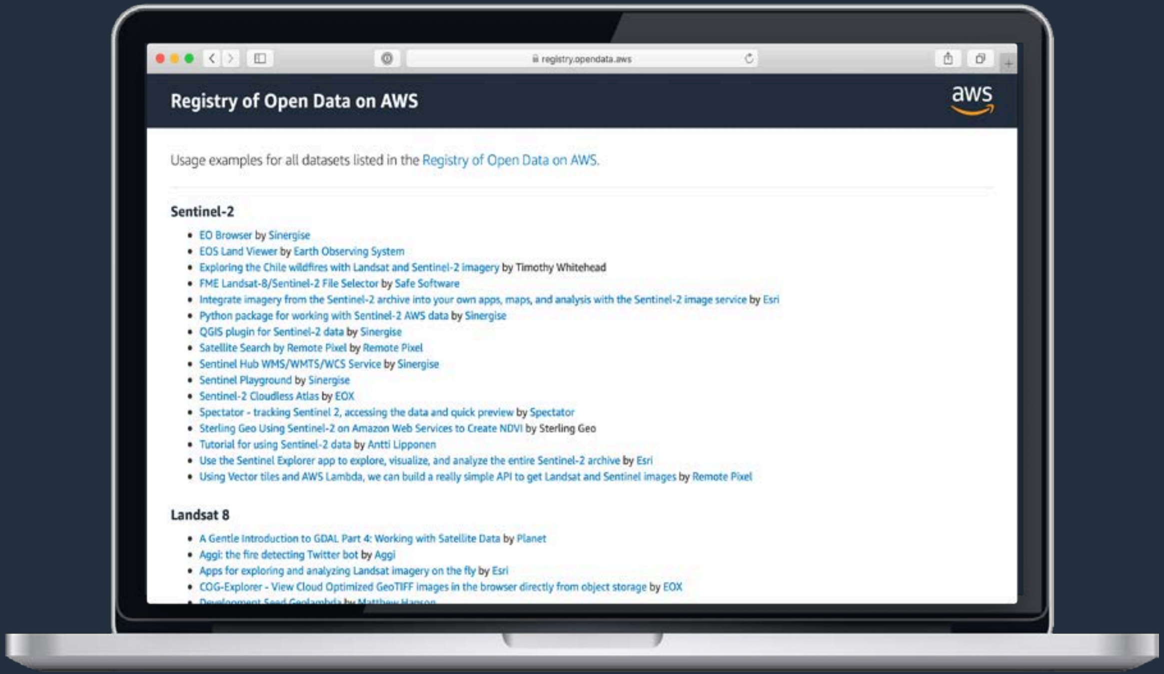
Move computing resources to the data.

NOAA datasets on AWS

- NEXRAD
- GOES 16 & 17
- Global Historical Climatology Network – Daily (GHCN-D)
- Global Historical Climatology Network – Hourly (GHCN-H)
- Global Ensemble Forecast System (GEFS)
- Global Forecast System (GFS)
- High-Resolution Rapid Refresh Model (HRRR)
- National Water Model Reanalysis & Short-Range Forecast
- Operational Forecast System (OFS)
- Integrated Surface Database (ISD)
- Global Hydro Estimator (GHE)

https://registry.opendata.aws/collab/noaa/

What does this enable?



https://registry.opendata.aws/usage-examples/

NCSU’s North Carolina Institute for Climate Studies

“We found that compared to a full-cost accounting of our current infrastructure, using AWS was much cheaper, and, with some guidance, our learning curve was relatively smooth and manageable. And the cloud can definitely be faster as there is almost no limit to the amount of parallel processing you can deploy, funds permitting.”

- ”Mistakes are cheap”
- “Costs are lower and more transparent”
- “Risks are low”

https://aws.amazon.com/blogs/publicsector/embracing-the-cloud-for-climate-research/

NEXRAD on AWS

- 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
- The Cornell Lab of Ornithology used the data on AWS to reveal 4 billion birds on the move

https://registry.opendata.aws/noaa-nexrad/

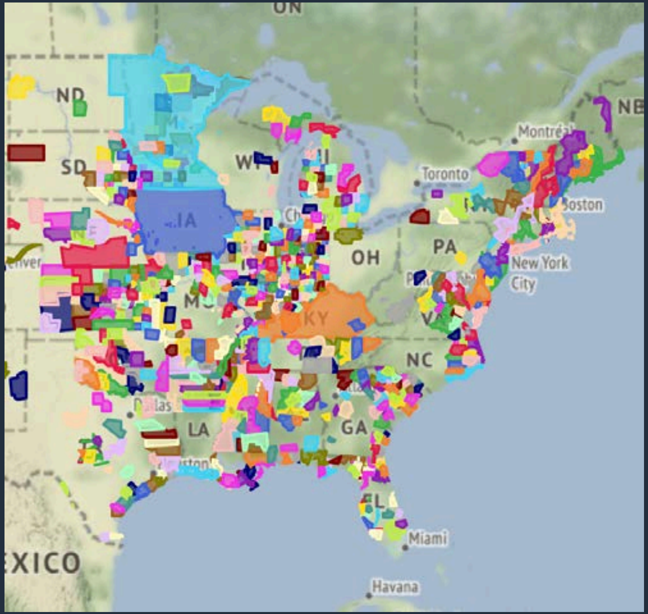
Filterable notification topics

{"Message" : {
  "S3Bucket": "unidata-nexrad-level2-chunks", "SiteID": "KDLH",
  "L2Version": "V06", "Key": "KDLH/500/20180202-052714",
  "VolumeID": 500, "ChunkType": "S", "ChunkID": 1
}, "MessageAttributes" : {
  "SiteID": {"Type":"String","Value":"KDLH"},
  "VolumeID”: {"Type":"Number","Value":"500"},
  "ChunkType": {"Type":"String","Value":"S”}
}}

Only be alerted for data that matters to you! (lower cost and complexity)

USGS 3DEP LiDAR point cloud

“The ability to use cloud computing with free, open 3DEP data will foster amazing new applications and uses of these data that we could not have done before. Just being able to see an entire statewide project with hundreds of billions of points at one time from a browser is amazing, let alone the potential to process and analyze all these data together in new and innovative ways.”



https://usgs.entwine.io/

NASA Cumulus: A case study

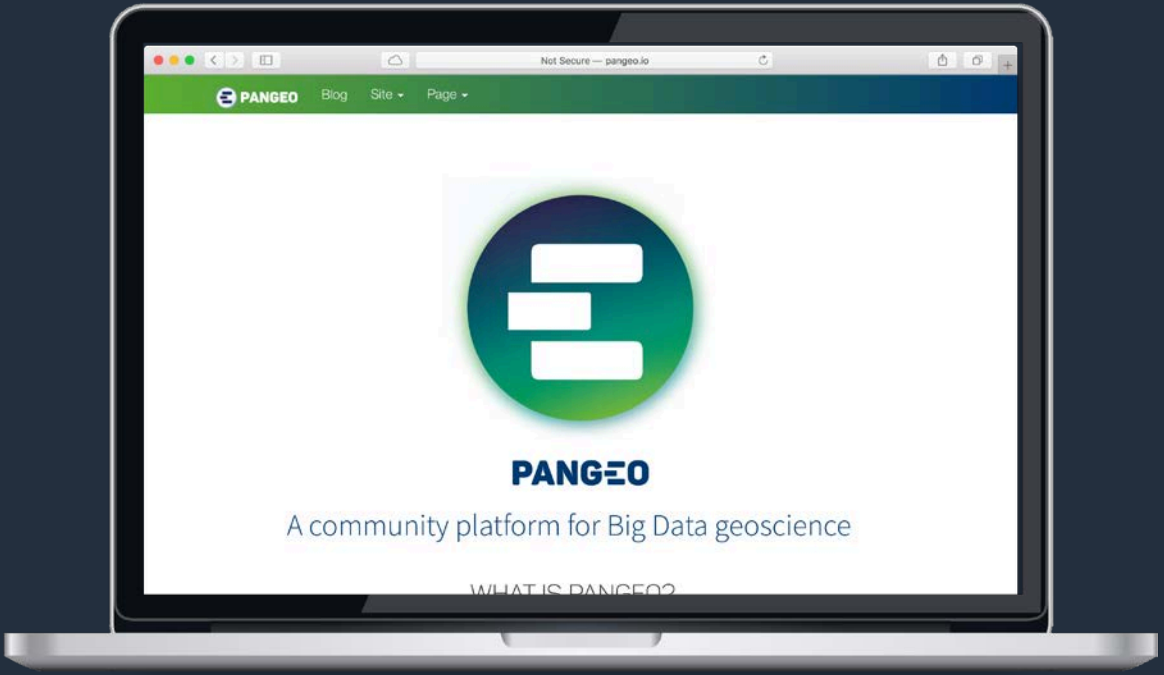
In 2016, NASA delivered more than 1.51 billion Earth science data products to more than 3 million data users around the world.

NASA's Earth Science Data Systems (ESDS) program is building a cloud-based platform to ingest, process, catalog, archive, and distribute NASA's Earth Data streams, expected to be 247 petabytes by 2025.

NASA Cumulus: Goals

- Data acquisition from data providers (such as NASA science teams)
- Data ingest (including validation and processing)
- The harvest, creation, and publication of dataset metadata to the EOSDIS Common Metadata Repository (CMR)
- The storage and distribution of data, including disaster recovery
- Publication of metrics to the ESDIS Metrics System (EMS), which collects and organizes various metrics from the DAACs and other data providers

Community efforts: Pangeo platform



http://pangeo.io

Community efforts: Spatio Temporal Asset Catalog (STAC)

The STAC specification aims to standardize the way geospatial assets are exposed online and queried. The initial focus is primarily remotely-sensed imagery (from satellites, but also planes, drones, balloons, etc), but the core is designed to be extensible to SAR, full motion video, point clouds, hyperspectral, LiDAR and derived data like NDVI, Digital Elevation Models, mosaics, etc.

- Static catalog
- Catalog API (https://www.element84.com/earth-search/)
- Core Metadata and extensions

https://github.com/radiantearth/stac-spec

Other thoughts

- Siloed datasets lead to siloed communities
    - Users of the NWM data on AWS uncovered an improperly documented flow rate

https://registry.opendata.aws/nwm-archive/

Sharing data (on AWS)

What we've learned

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

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

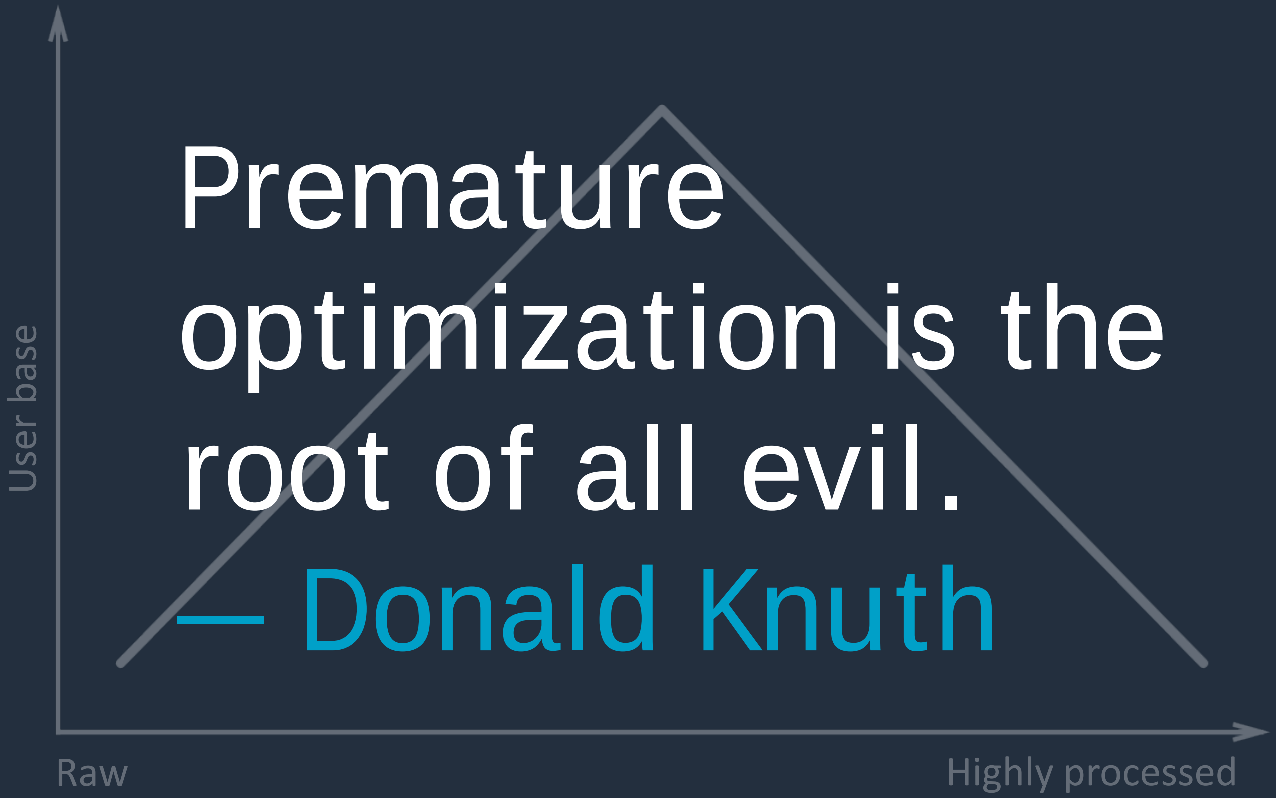
Amazon Resource Name (ARN)
`arn:aws:s3:::commoncrawl`

AWS Region
`us-east-1`

What makes a dataset successful?

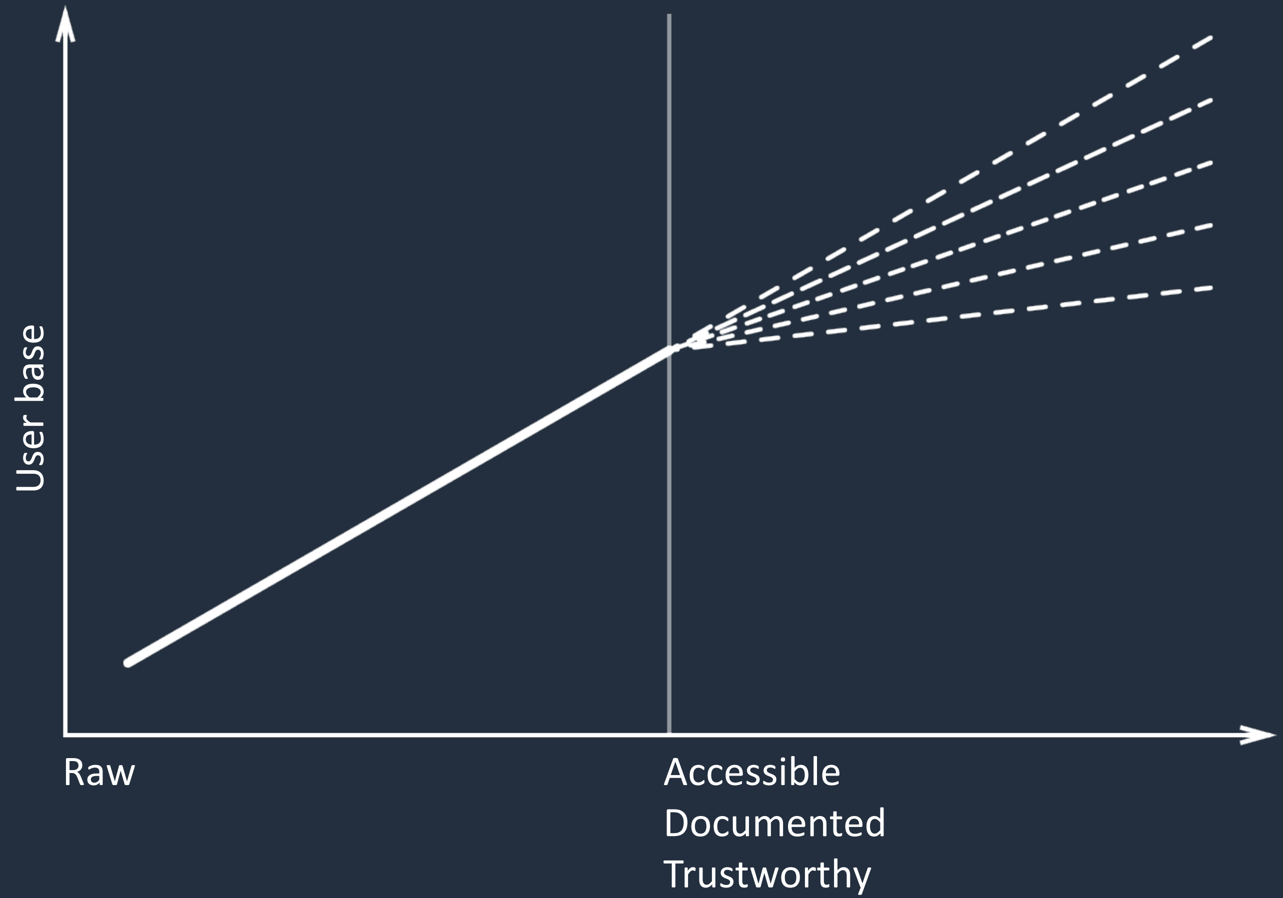
It is treated like a product.
It is optimized for analysis.


User base
Raw
Highly processed

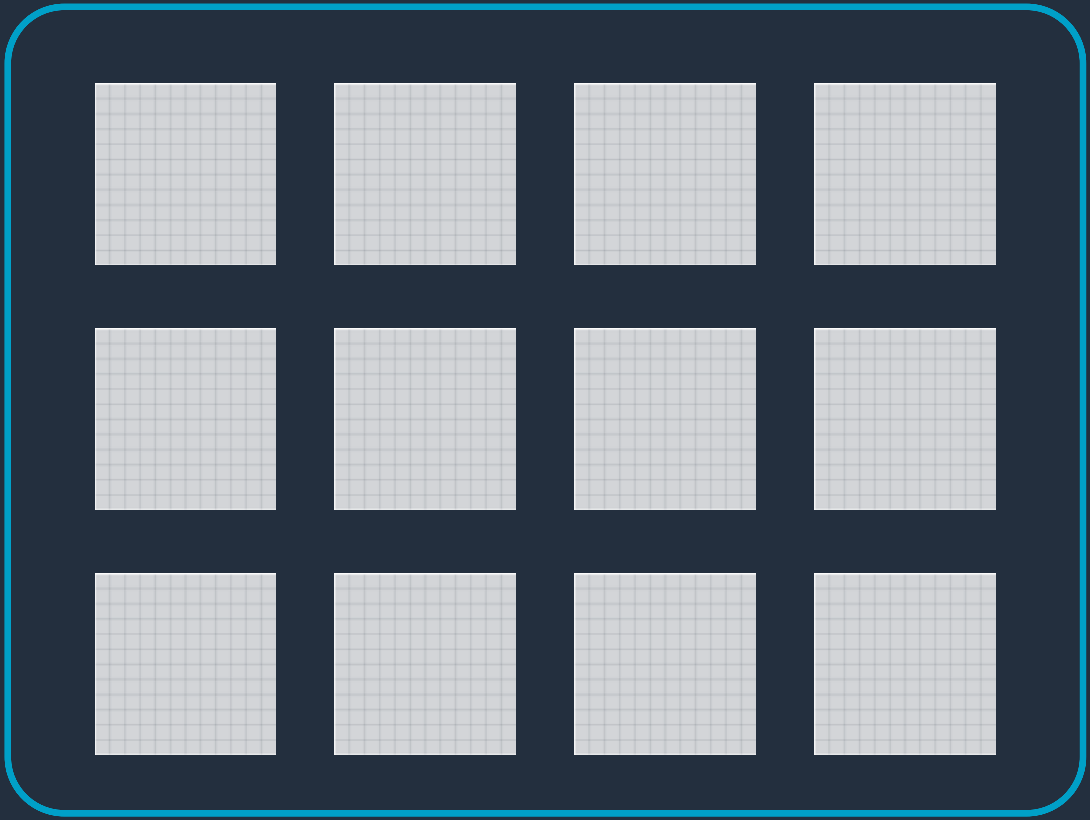


Premature optimization is the root of all evil.

— Donald Knuth

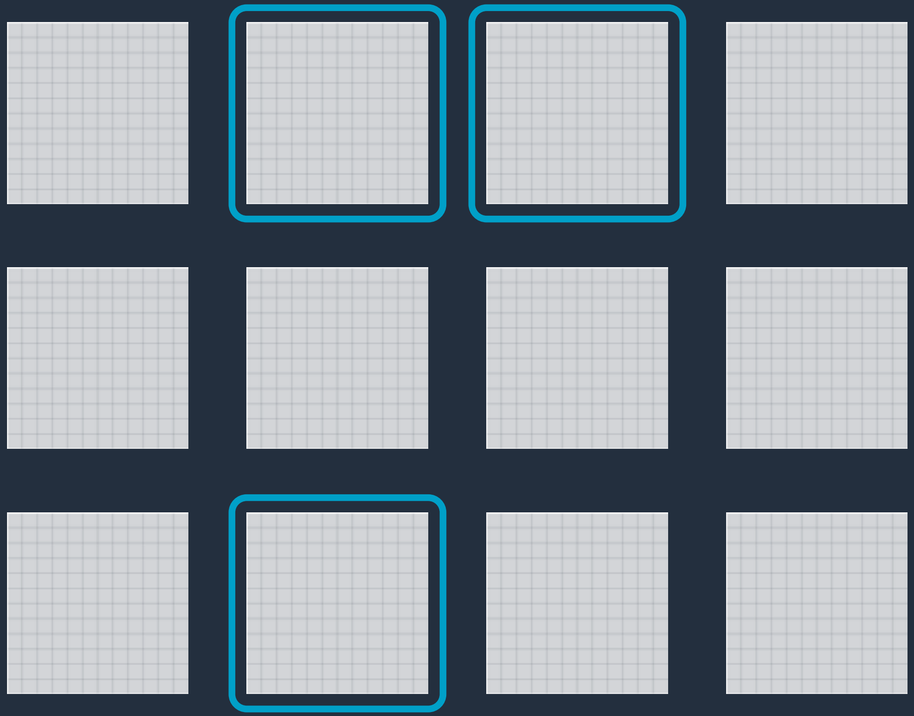

User base
Raw
Accessible
Documented
Trustworthy

The cloud-optimized GeoTIFF

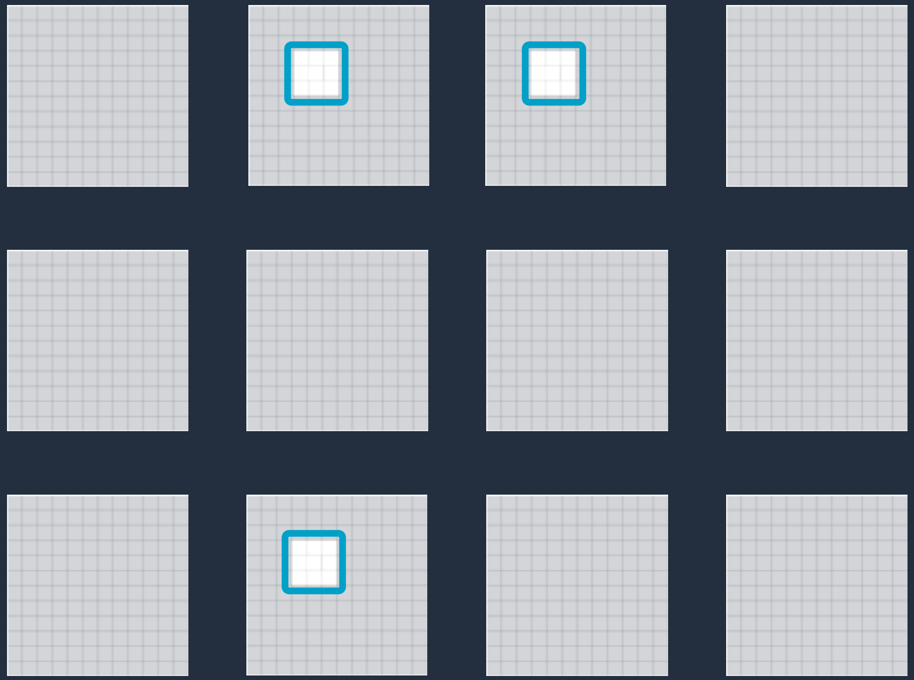


.tar

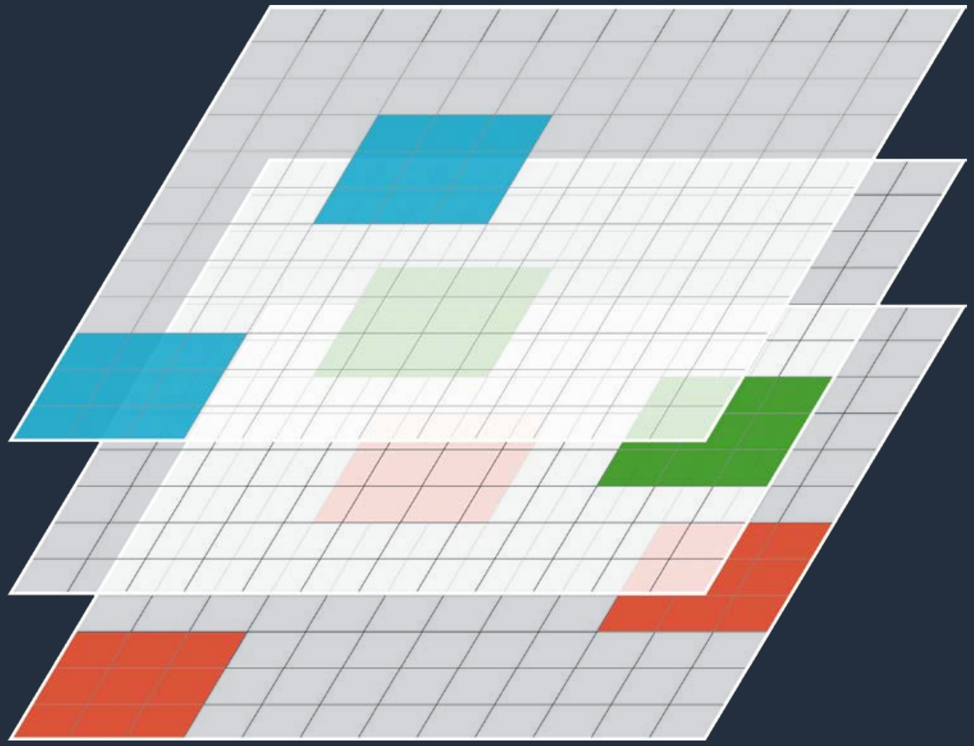
The cloud-optimized GeoTIFF



The cloud-optimized GeoTIFF

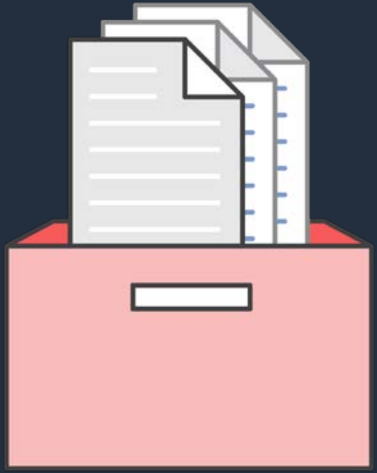


The cloud-optimized GeoTIFF



Indexing 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

https://registry.opendata.aws/noaa-goes/

Example: IRS 990 CSV as external index

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

What makes a dataset successful?

It is treated like a product.
It is optimized for analysis.
There is a community around it.

Thank you!

jflasher@amazon.com