

AWS
re:Invent

W P S 3 0 6

AWS Public Datasets: Lessons from staging petabytes of data for analysis

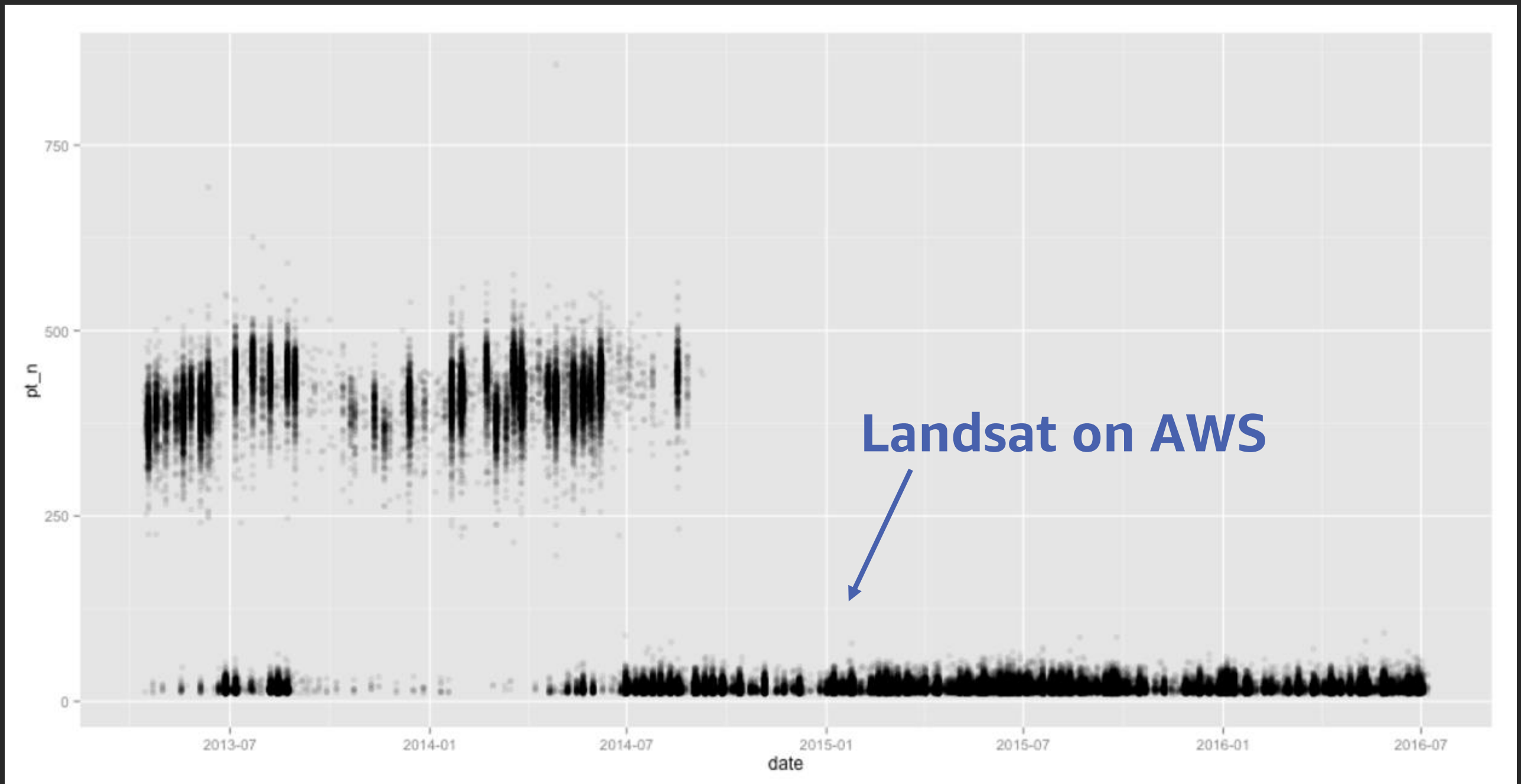
Jed Sundwall

Global Open Data Lead
Amazon Web Services



Landsat on AWS

<https://registry.opendata.aws/landsat-8/>



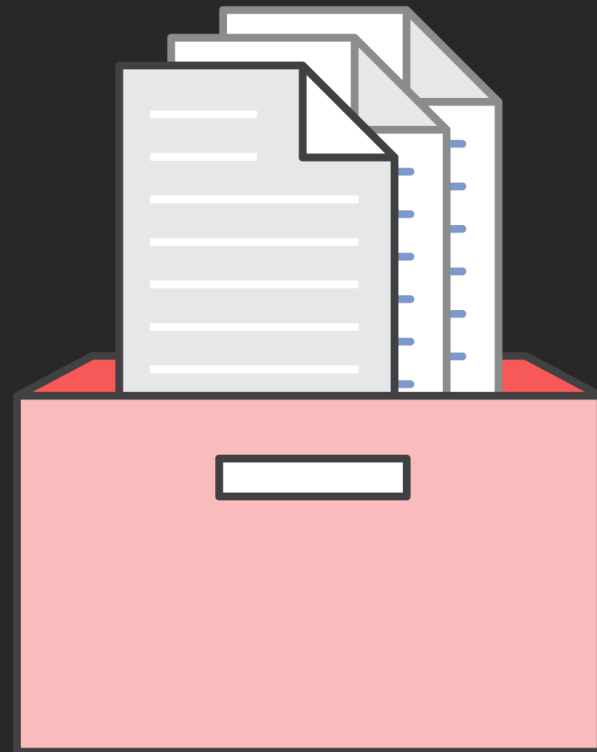
Graph by Drew Bollinger (@drewbo19) at Development Seed

Using serverless to visualize and analyze imagery

The screenshot displays the EO Browser interface. The top navigation bar includes a search bar, a 'Visualization' tab, and a 'My pins' section. The main content area shows satellite data for SENTINEL-2 L1C on 2018-03-26. A sidebar on the left provides various visualization options: Custom, True color, False color, False color (urban), NDVI, Moisture Index, SWIR, and NDWI. The main map area shows a false-color satellite image of a landscape with a circular selection and a path of four white dots. The bottom of the interface features a scale bar (300 m), a footer with 'Powered by SinerGise with contributions from the European Space Agency', and a status bar with coordinates (Lat: 42.07867, Lng: -5.56153) and map data attribution (Carto © CC BY 3.0, OpenStreetMap © ODbL).

Patterns

Amazon S3
key index



External index



Internal index



Allen Brain Observatory key naming index

registry.opendata.aws/allen-brain-observatory

```
visual-coding-2p
├── manifest.json          # used by AllenSDK to look up file paths
├── experiment_containers.csv # metadata for each container (area, imaging depth,
├── ophys_experiments.csv  # metadata for each experiment session
├── ophys_experiment_data  # traces, running speed, etc per experiment session
│   ├── <experiment_id>.nwb
│   └── ...
├── ophys_experiment_analysis # analysis files per experiment session
│   ├── <experiment_id>_<session_name>.h5
│   └── ...
└── ophys_movies           # motion-corrected video per experiment session
    ├── ophys_experiment_<experiment_id>.h5
    └── ...
```

Landsat imagery naming index

registry.opendata.aws/landsat-8

Accessing Landsat on AWS

The data are organized using a directory structure based on each scene's path and row. For instance, the files for Landsat scene LC08_L1TP_139045_20170304_20170316_01_T1 are available in the following location:

s3://landsat-pds/c1/L8/139/045/LC08_L1TP_139045_20170304_20170316_01_T1/

The "c1" refers to Collection 1, the "L8" refers to Landsat 8, "139" refers to the scene's path, "045" refers to the scene's row, and the final directory matches the product's identifier, which uses the following naming convention:

`LXSS_LLLL_PPPRRR_YYYYMMDD_yyyymmdd_CC_TX`, in which:

- `L` = Landsat
- `X` = Sensor
- `SS` = Satellite
- `PPP` = WRS path
- `RRR` = WRS row
- `YYYYMMDD` = Acquisition date
- `yyymmdd` = Processing date
- `CC` = Collection number
- `TX` = Collection category

In this case, the scene corresponds to WRS path 139, WRS row 045, and was taken on March 4th, 2017.

IRS 990 external index – CSV

registry.opendata.aws/irs990

	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 FOUNDATION	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 NETW	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

Common Crawl external index – Parquet

registry.opendata.aws/commoncrawl

```
1 SELECT COUNT(*) AS count,  
2     url_host_registered_domain  
3 FROM "ccindex"."ccindex"  
4 WHERE crawl = 'CC-MAIN-2018-05'  
5     AND subset = 'warc'  
6     AND url_host_tld = 'no'  
7 GROUP BY url_host_registered_domain  
8 HAVING (COUNT(*) >= 100)  
9 ORDER BY count DESC
```

Run Query

Save As

Format query

New Query

(Run time: 5.38 seconds, Data scanned: 2.12MB)

Results

	count	url_host_registered_domain
1	278171	blogg.no
2	177143	aftenposten.no
3	172755	blogspot.no

What makes a dataset successful?

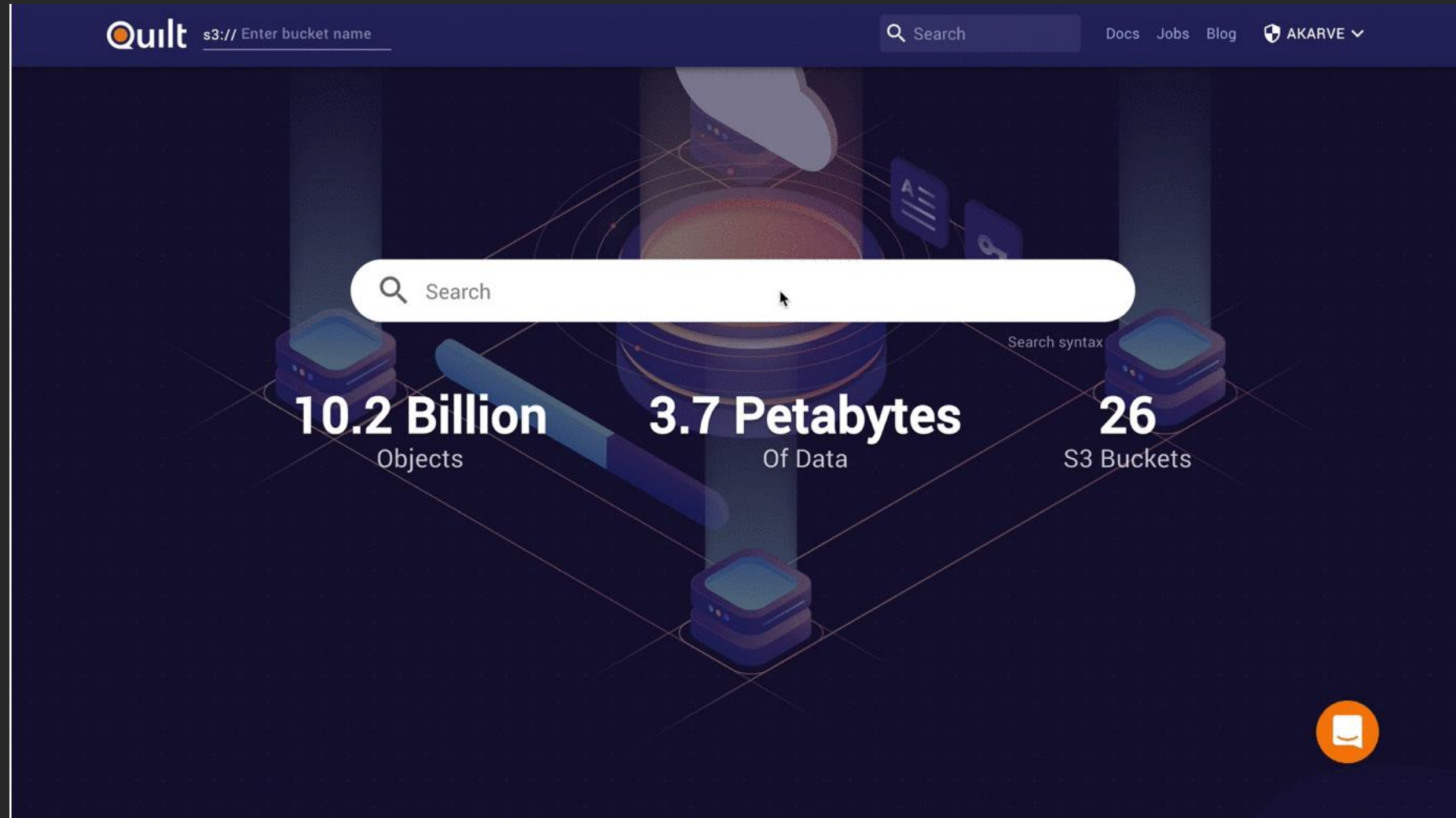
It is treated like a product.

It is optimized for analysis.

There is a community around it.

Data discovery with Elastic Search

open.quiltdata.com



The screenshot shows the Quilt website interface. At the top left is the Quilt logo and a text input field with the placeholder "s3:// Enter bucket name". To the right is a search bar with a magnifying glass icon and the text "Search". Further right are links for "Docs", "Jobs", "Blog", and a user profile icon labeled "AKARVE" with a dropdown arrow. The main content area features a large, stylized illustration of a hand holding a glowing orb, surrounded by server racks and data blocks. A prominent white search bar is overlaid on this illustration, containing a magnifying glass icon and the text "Search". Below the search bar, three statistics are displayed: "10.2 Billion Objects", "3.7 Petabytes Of Data", and "26 S3 Buckets". A "Search syntax" link is visible to the right of the search bar. In the bottom right corner, there is an orange circular icon with a white speech bubble.

Quilt s3:// Enter bucket name

Search

Docs Jobs Blog AKARVE

Search

Search syntax

10.2 Billion
Objects

3.7 Petabytes
Of Data

26
S3 Buckets

Data documentation with AWS Lambda, Amazon API Gateway

open.quiltdata.com/b/allencell

The screenshot shows the Quilt data catalog interface for the 'allencell' bucket. The top navigation bar includes the Quilt logo, the bucket path 's3://allencell', a search bar, and links for 'Docs', 'Jobs', 'Blog', and 'AKARVE'. Below the navigation, there are tabs for 'OVERVIEW', 'FILES', and 'PACKAGES'. The main content area features a dark blue header with the bucket name 'allencell' and a descriptive paragraph. Below this, statistics show '462k Objects' and '7.7 TB' of data, with a note that the data is from the 'Registry of Open Data on AWS'. Two charts are displayed: 'Objects by File Extension' and 'Downloads (total): 200.1k' for the 'LAST 1 MONTH'. The 'Downloads' chart is a stacked area chart showing growth over time. At the bottom, there is a section for 'Allen Cell Imaging Collections' with a 'Description' sub-section and a Quilt logo icon.

allencell

This bucket contains multiple datasets (as Quilt packages) created by the Allen Institute for Cell Science (AICS). The imaging data in this bucket contains either of the following: 1) field of view images from glass plates 2) cell membrane, DNA, and structure segmentations 3) cell membrane, DNA and structure contours 4) machine learning imaging predictions of the previously listed modalities. In addition, many of the datasets include CSVs that contain feature sets related to that data.

462k Objects **7.7 TB**

From the [Registry of Open Data on AWS](#)

Objects by File Extension

File Extension	Size	Count
other	4.2 TB	99.2k
.tiff	3.5 TB	127.6k
.png	2.4 GB	156.8k
.json	440.5 MB	78.4k
.csv	200.2 MB	5
.md	61.1 kB	8

Downloads (total): 200.1k

LAST 1 MONTH

Allen Cell Imaging Collections

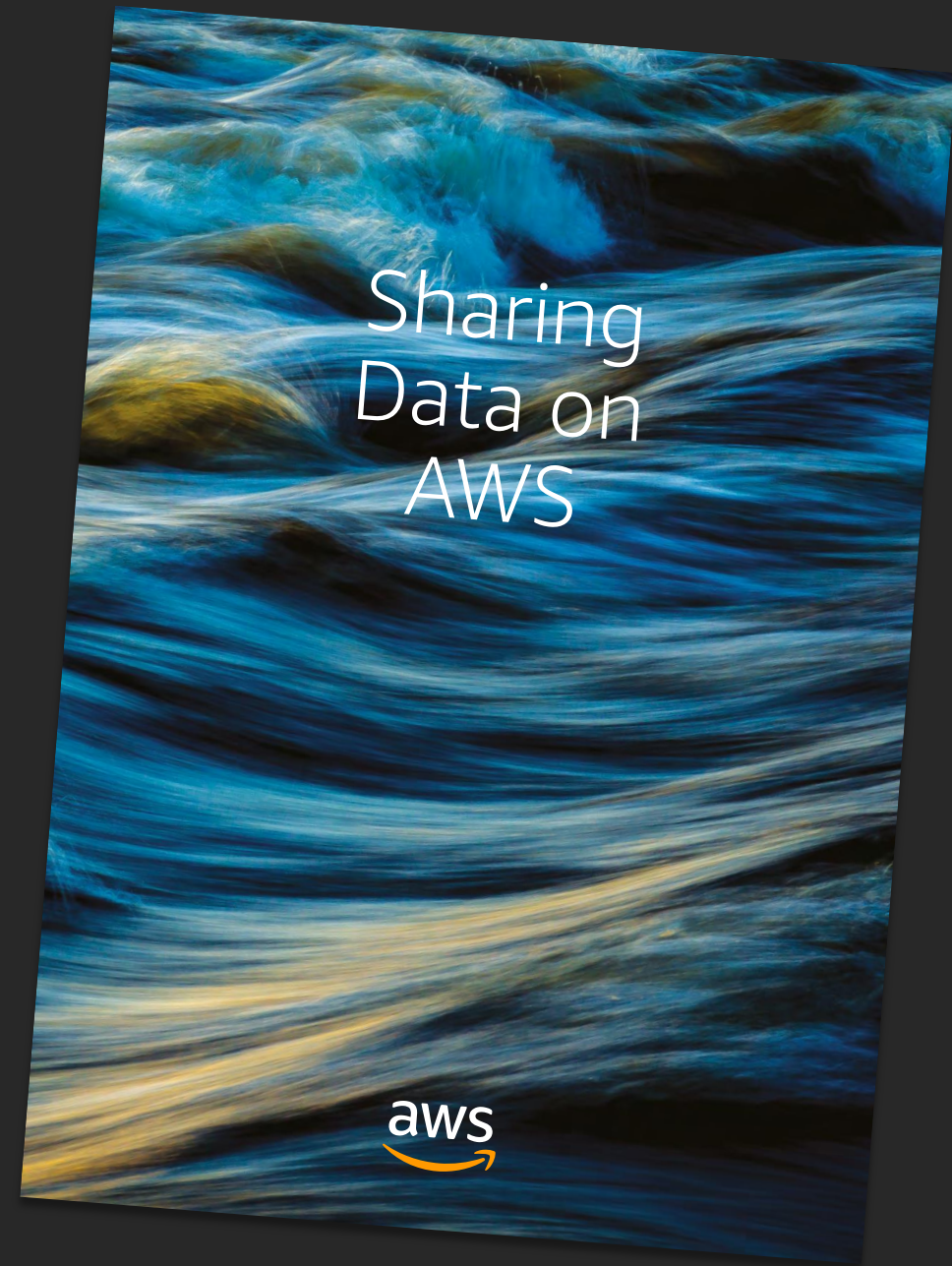
Description

Guide to sharing data on AWS

Over 40 pages of insights on data sharing and case studies from customers, including:

- Transport for London
- Canada's Communications Security Establishment
- The US Geological Survey
- The Allen Institute for Brain Science

opendata.aws/guide-pdf



Thank you!

Jed Sundwall

jed@amazon.com

Aneesh Karve

aneesh@quiltdata.io



Please complete the session survey in the mobile app.