

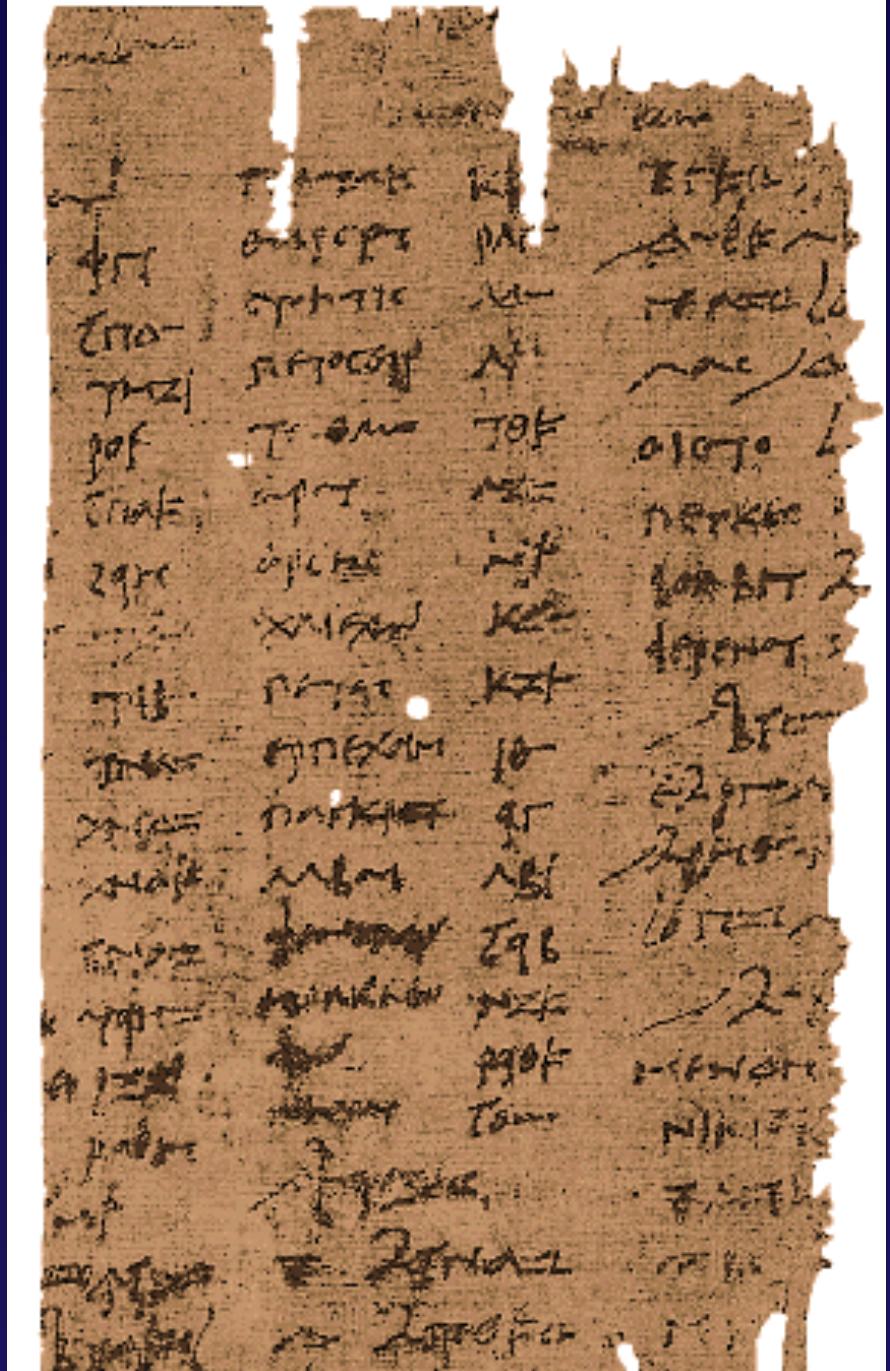


Streaming Data Collection Concepts



Brock Tubre

INSTRUCTOR



age	village	town
0-4	24	26
5-9	16	26
10-14	7	16
15-19	13	23
20-24	13	18
25-29	14	12
30-34	16	19
35-39	10	6
40-44	12	6
45-49	13	11
50-54	4	6
55-59	11	3
60-64	3	3
65-69	5	2
70-74	6	3
75-79	2	0
80-84	0	1

Early Data Collection

Census of Egypt

The practice of collecting census data began in Egypt second millennium BC (2000 through 1001 BC), where it was used for tax gathering and to determine fitness for military services.

Where does data come from?

WHERE DOES DATA COME FROM?



Datasets | Kaggle

https://www.kaggle.com/datasets

kaggle Search Competitions Datasets Kernels Discussion Learn ...

Datasets

Documentation New Dataset

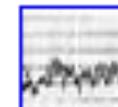
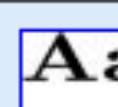
Public Your Datasets Favorites Sort by Hotness

14,103 Datasets Sizes File types Licenses Tags Search datasets

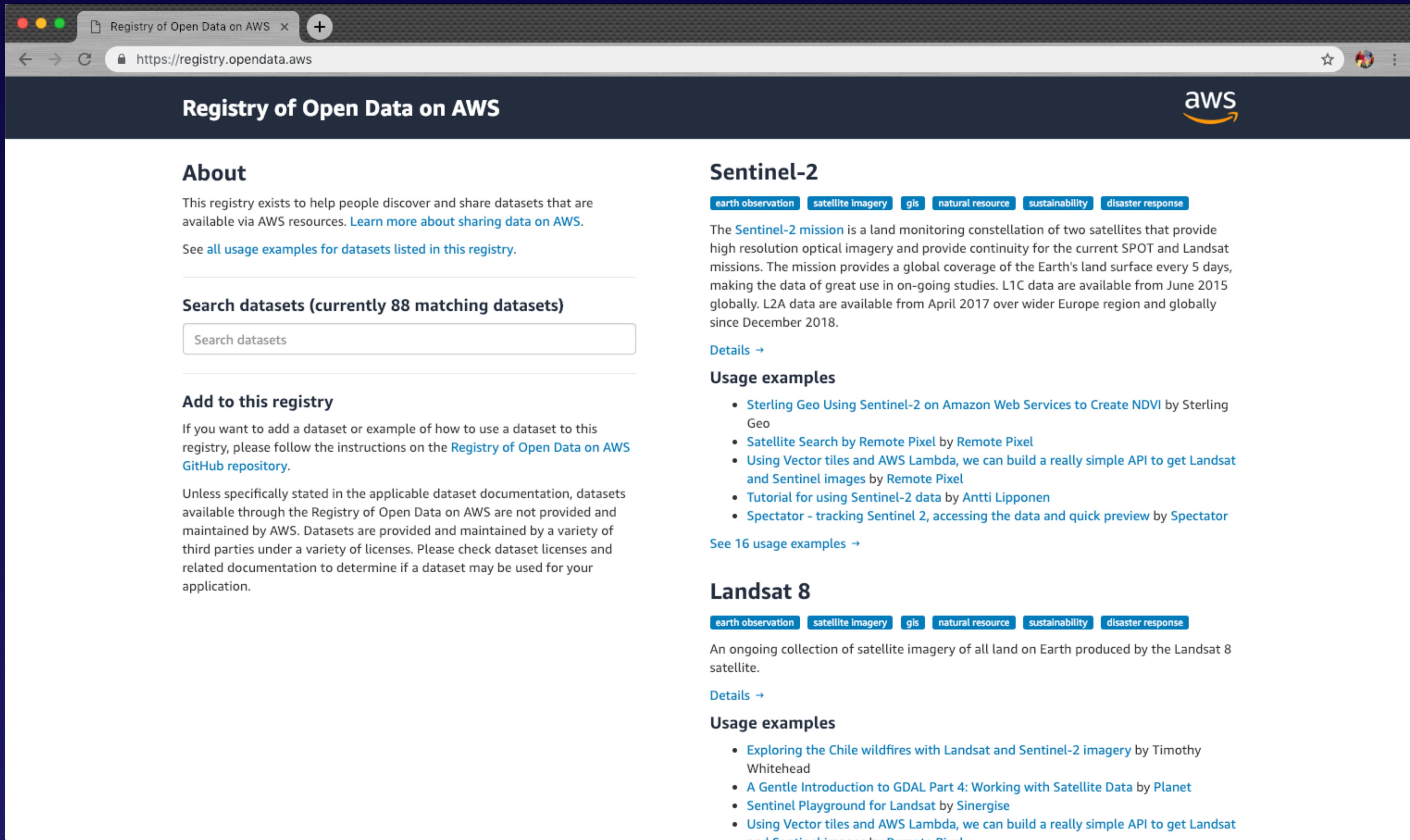
	Graduate Admissions Predicting admission from important parameters Mohan S Acharya updated 21 days ago (Version 2)	regression ... model com... random for... + 2 more...	CSV 9.4 KB CC0	</> 44 6 18k
	FiveThirtyEight Comic Characters Dataset Explore Data from FiveThirtyEight FiveThirtyEight Maintained by Kaggle updated 4 days ago (Version 107)	CSV 577.1 KB CC0	</> 3 0 2k	
	FIFA 19 complete player dataset 18k+ FIFA 19 players, 85+ attributes extracted from the latest FIFA database Karan Gadiya updated a month ago	video games american f... association... + 2 more...	CSV 2.1 MB CC4	</> 17 3 15k
	Amazon Alexa Reviews A list of 3150 Amazon customers reviews for Alexa Echo, Firestick, Echo Dot etc. Manu Siddhartha updated 6 months ago (Version 3)	beginner nlp deep learni... + 2 more...	Other 163.6 KB Other	</> 30 1 15k
	Kuzushiji-MNIST Classify handwritten characters in ancient Japanese manuscripts anokas updated a month ago (Version 3)	literature classification image data + 2 more...	Other 318.1 MB CC4	</> 11 0 3k
	Google Play Store Apps	video games computer s...	CSV 19 MR	</> 174 20

UCI Machine Learning Repository

Browse Through: **22 Data Sets** [Table View](#) [List View](#)

Default Task	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Classification (19)	 Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Regression (3)	 Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Clustering (0)	 Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Other (1)	 Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
Attribute Type	 Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
Categorical (8)	 Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
Numerical (3)	 Audiology (Original)	Multivariate	Classification	Categorical	226		1987
Mixed (10)							
Data Type							
Multivariate (20)							
Univariate (1)							
Sequential (0)							
Time-Series (0)							
Text (1)							
Domain-Theory (0)							
Other (2)							
Area							
Life Sciences (8)							
Physical Sciences (1)							
CS / Engineering (2)							
Social Sciences (4)							
Business (0)							
Game (2)							
Other (5)							
# Attributes							

mlr.cs.umass.edu/ml/datasets.html



The screenshot shows the Registry of Open Data on AWS website. The top navigation bar includes the AWS logo and a search bar. The main content area displays two dataset entries: "Sentinel-2" and "Landsat 8". Each entry includes a brief description, a "Details" link, and a "Usage examples" section with a list of links.

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](#).

Search datasets (currently 88 matching datasets)

Search datasets

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

- [Sterling Geo Using Sentinel-2 on Amazon Web Services to Create NDVI](#) by Sterling Geo
- [Satellite Search by Remote Pixel](#) by Remote Pixel
- [Using Vector tiles and AWS Lambda, we can build a really simple API to get Landsat and Sentinel images](#) by Remote Pixel
- [Tutorial for using Sentinel-2 data](#) by Antti Lipponen
- [Spectator - tracking Sentinel 2, accessing the data and quick preview](#) by Spectator

[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

- [Exploring the Chile wildfires with Landsat and Sentinel-2 imagery](#) by Timothy Whitehead
- [A Gentle Introduction to GDAL Part 4: Working with Satellite Data](#) by Planet
- [Sentinel Playground for Landsat](#) by Sinergise
- [Using Vector tiles and AWS Lambda, we can build a really simple API to get Landsat and Sentinel images](#) by Remote Pixel

Google's Big Query

BigQuery - demo-dms-project

https://console.cloud.google.com/bigquery?_ga=2.241429698.-1185341951.1542652994&project=focus-flight-222112&folder&organizationId

Google Cloud Platform demo-dms-project

BigQuery BETA Go to Classic UI + COMPOSE NEW QUERY

Query history

Saved queries

Job history

Transfers

Resources + ADD DATA

Search for your tables and datasets

focus-flight-222112

bigquery-public-data

- austin_311
- austin_bikeshare
- austin_crime
- austin_incidents
- austin_waste
- baseball
- bitcoin_blockchain
- bls
- census_bureau_construction
- census_bureau_international
- census_bureau_usa
- census_fips_codes
- chicago_crime
- chicago_taxi_trips
- cloud_storage_gcs_index

Unsaved query Edited

```
1 SELECT * FROM `fh-bigquery.reddit_posts.2017_09` WHERE subreddit IN ('aws', 'googlecloud', 'AZURE');
```

Run Save query Save view More This query will process 4.72 GB when run.

Query history REFRESH

Sort by Date Filter queries

11/21/18

Query canceled

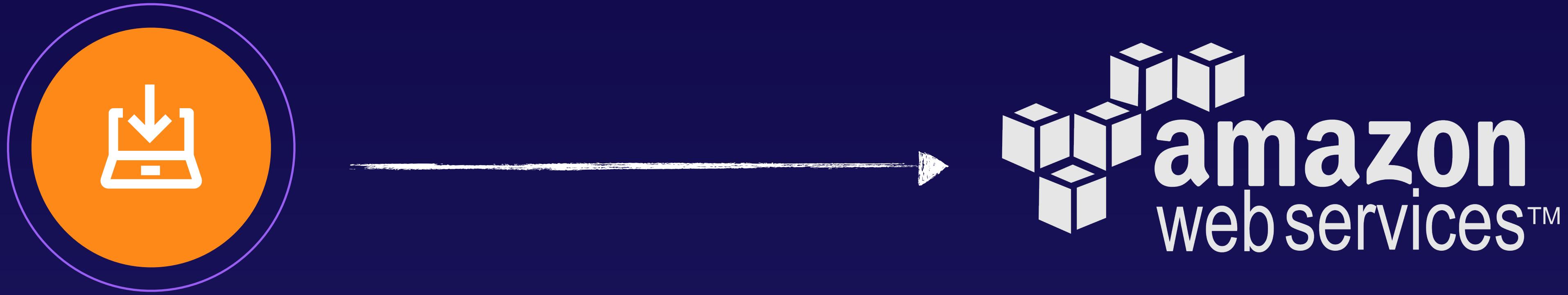
Query canceled 7:48 PM

Open query in editor

```
1 SELECT * FROM `fh-bigquery.reddit_posts.2017_09` WHERE subreddit IN ('aws', 'googlecloud', 'AZURE');
```

Job ID	focus-flight-222112:US.bquxjob_7a1e784e_16738e35b21
User	brocktubre@gmail.com
Location	United States (US)
Creation time	Nov 21, 2018, 7:48:14 PM
Start time	Nov 21, 2018, 7:48:14 PM
End time	Nov 21, 2018, 7:49:24 PM
Duration	1 min 9.777 sec

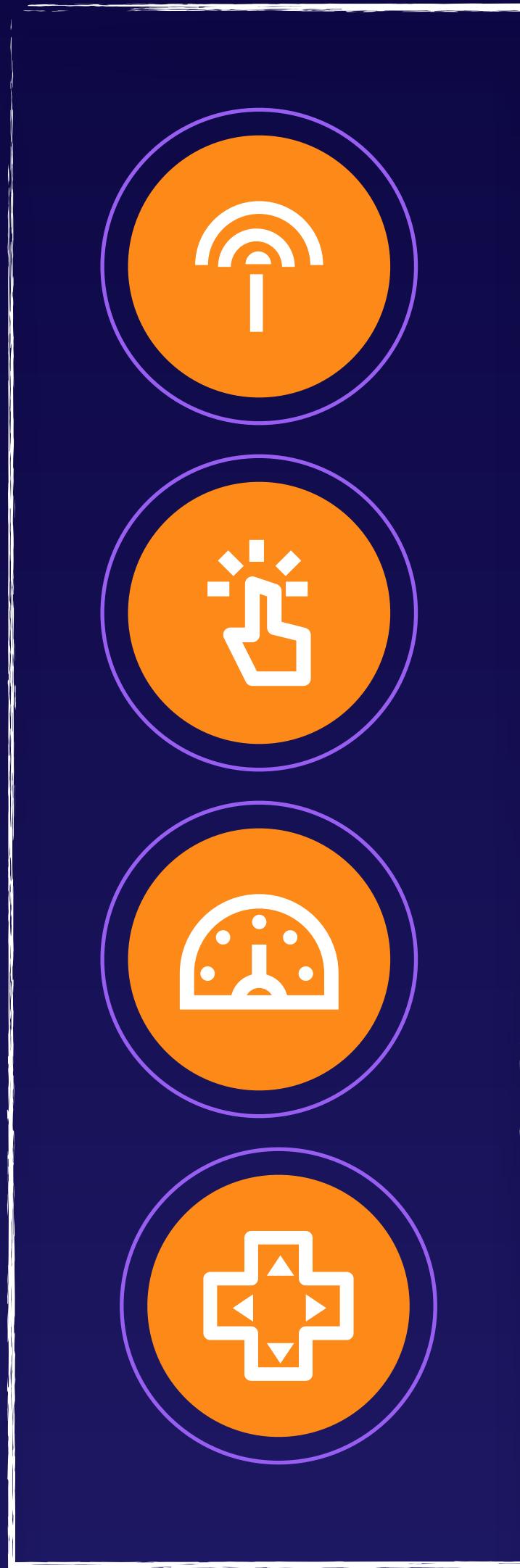
Load Data into AWS



What about streaming data?

WHAT ABOUT STREAMING DATA?

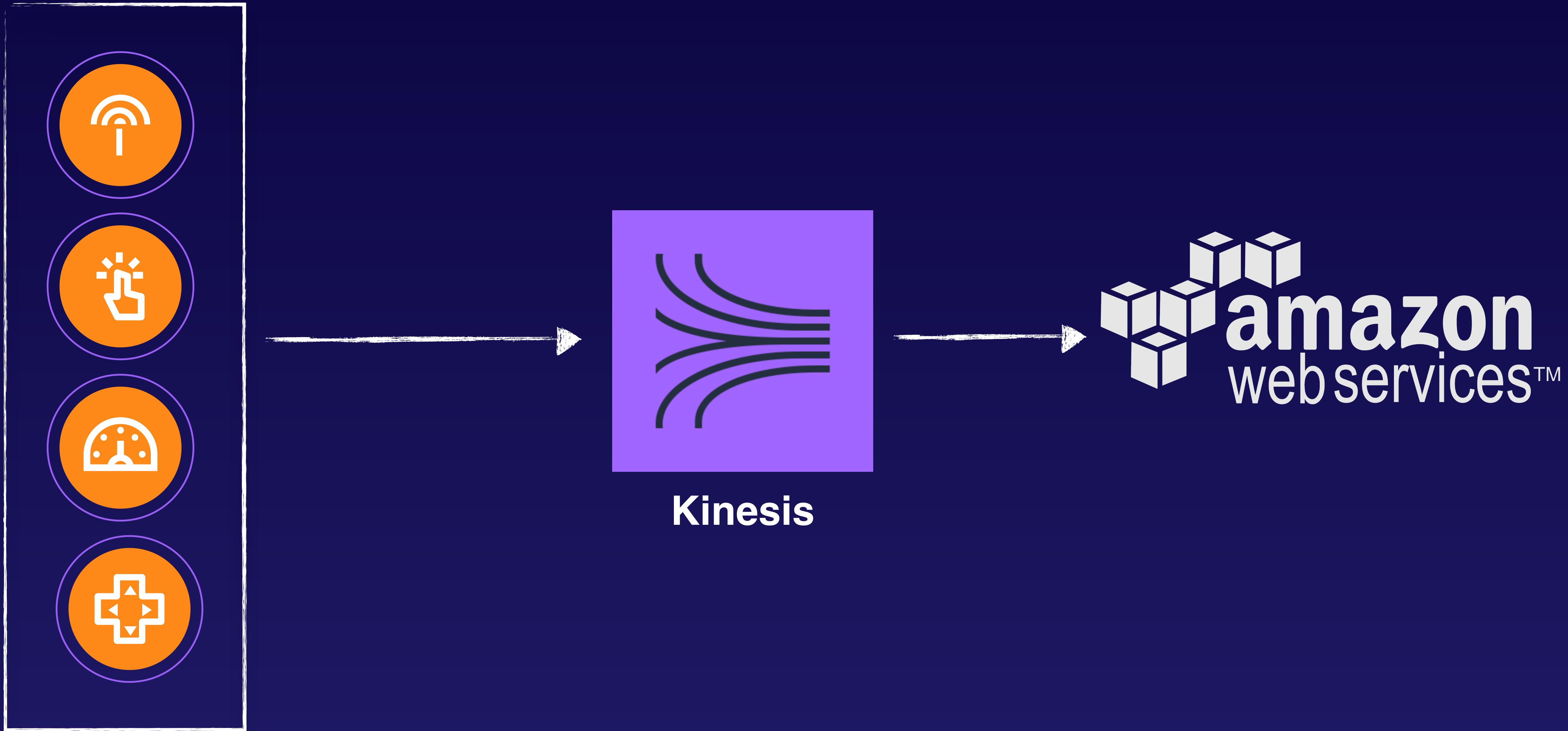






?





The Marvel Family

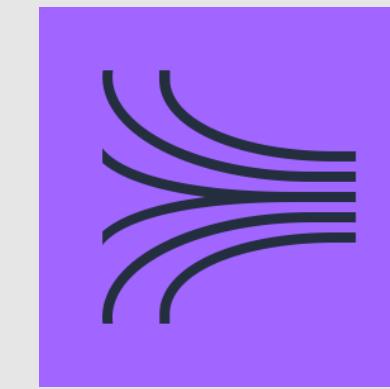
“The Shazam Family”

Family of superheroes created in 1942 including characters such as Captain Marvel “Shazam”, Mary Marvel, Captain Marvel Jr., Uncle Marvel.

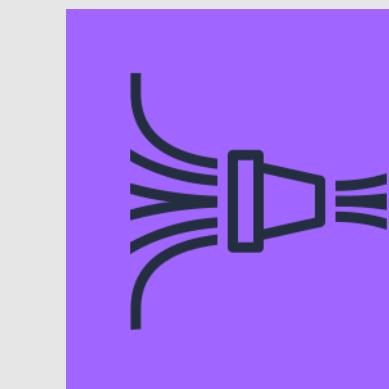
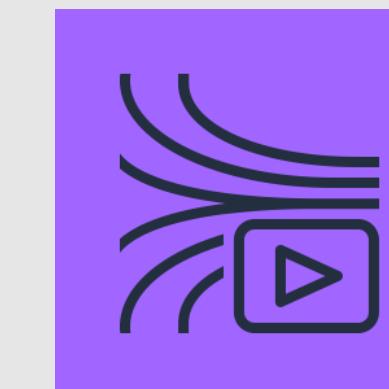
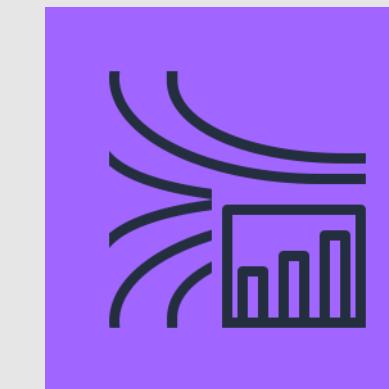


The Kinesis Family

The Kinesis Family



Kinesis

Kinesis
Data StreamsKinesis
Data FirehoseKinesis
Video StreamsKinesis
Data Analytics