

Scaling biodiversity data processing and analysis workflows with Apache Beam



Jeremy Malczyk
Software Engineer
Map of Life, Yale University

Walter Jetz (PI, Yale University)
Robert Guralnick (Co-I, University of Florida, Gainesville)
Adam Wilson (Co-I, State University of New York, Buffalo)
Map of Life software development team (<https://mol.org/team>)

Yale



Software Workflows and Tools for Integrating Remote Sensing and Organismal Occurrence Data Streams to Assess and Monitor Biodiversity Change

2016 ROSES A.41 Solicitation NNH16ZDA001N
AIST Research Opportunities in Space and Earth Sciences
Grant # AIST-16-0092

Walter Jetz (PI, Yale University)

Robert Guralnick (Co-I, University of Florida, Gainesville)

Adam Wilson (Co-I, State University of New York, Buffalo)

WHY?

“improving the ease with which the biology and ecology communities can understand, select and use appropriately NASA remote sensing data.”

- Types of biodiversity data
- Combining biodiversity data with remote sensing products
- Existing tools for data fusion
- Scaling with Beam

Types of biodiversity data

- Observation / occurrence
- Expert range maps
- Local Inventories
- Gridded surveys
- Regional checklists
- Distribution model predictions

MOL
MAP OF LIFE

login/register en de es fr zh

Species Locations Indicators Patterns

Species Home Summary Map Detailed Map Search for a species

White Stork

Ciconia ciconia

Storks

Least Concern (IUCN 2016)

Sources

- Point observations 300,600 Q i
- Regional checklists 2,562 Q i
- Local inventories 47 Q i
- Expert range maps 2 Q i
 - Jetz et al. (2012) Q i
 - MOL grid of BirdLife Q i
- Gridded surveys 210 Q i

Map Satellite

Google

Map data ©2018 Google, INEGI Terms of Use

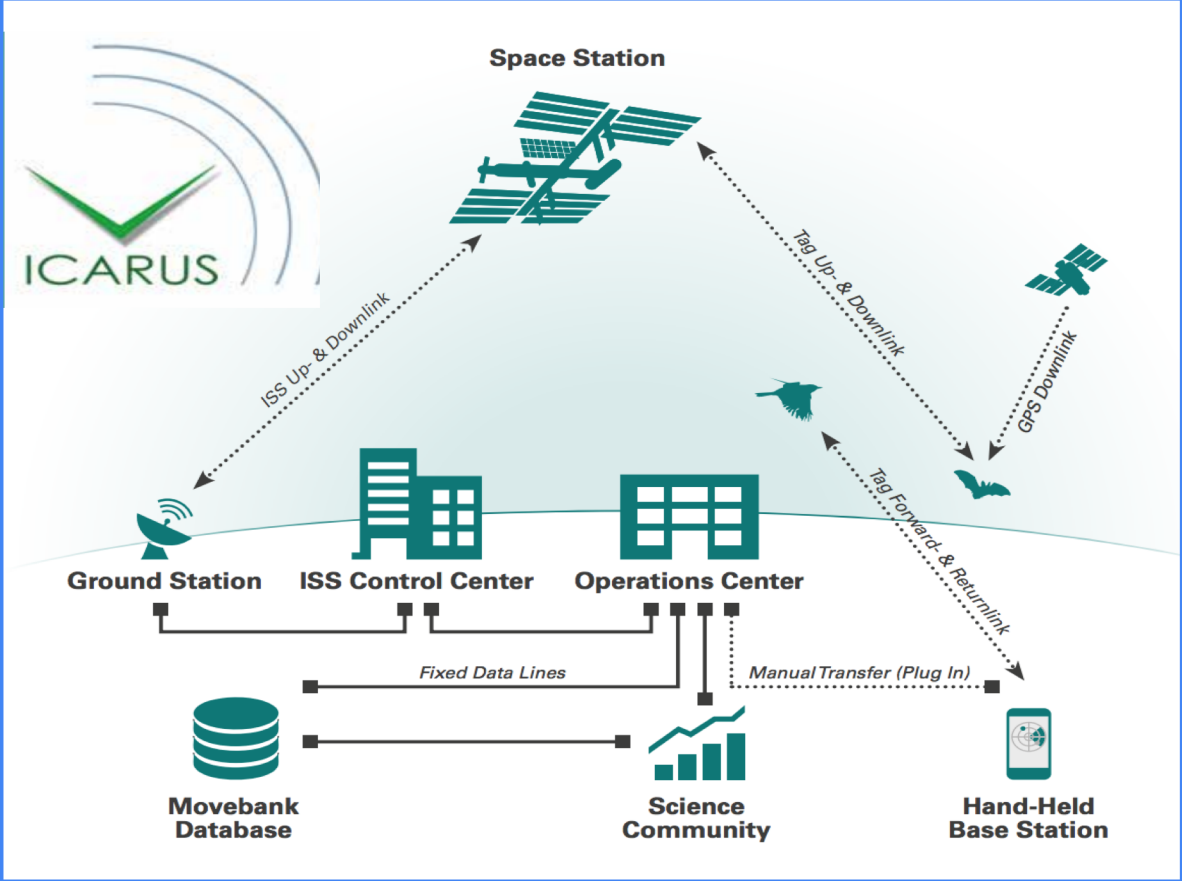
supporters feedback privacy policy terms of service

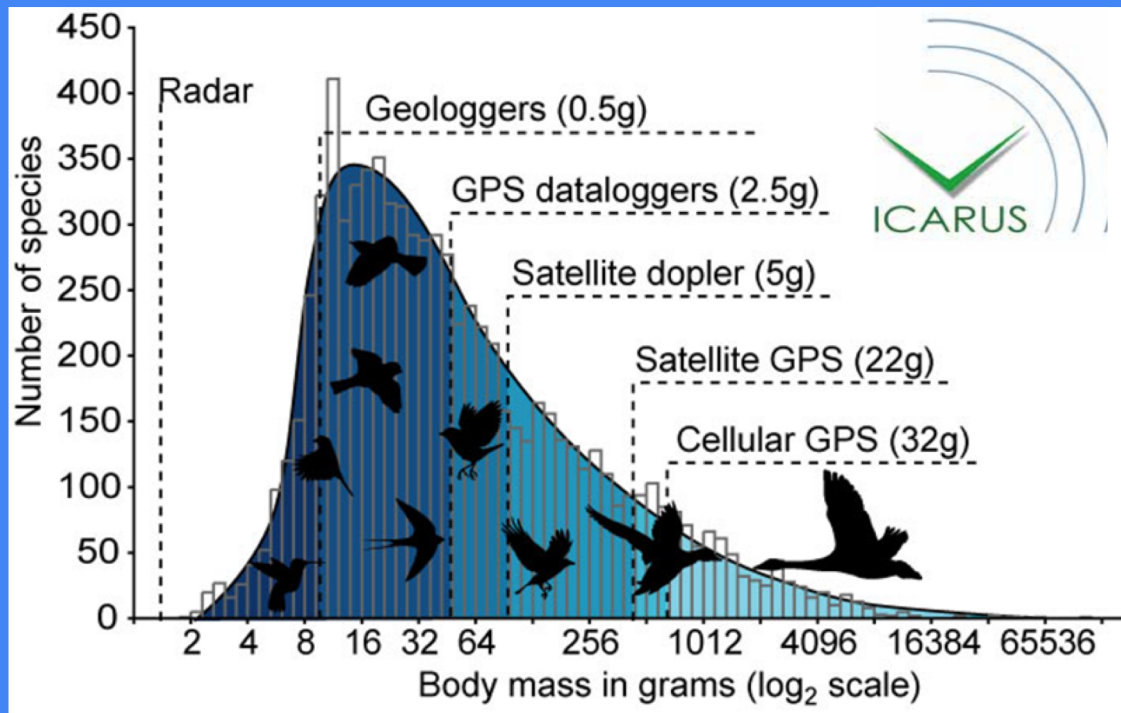
Movement data

The screenshot displays the Movebank website interface. At the top, the Movebank logo and navigation menu are visible. The main content area is divided into three sections:

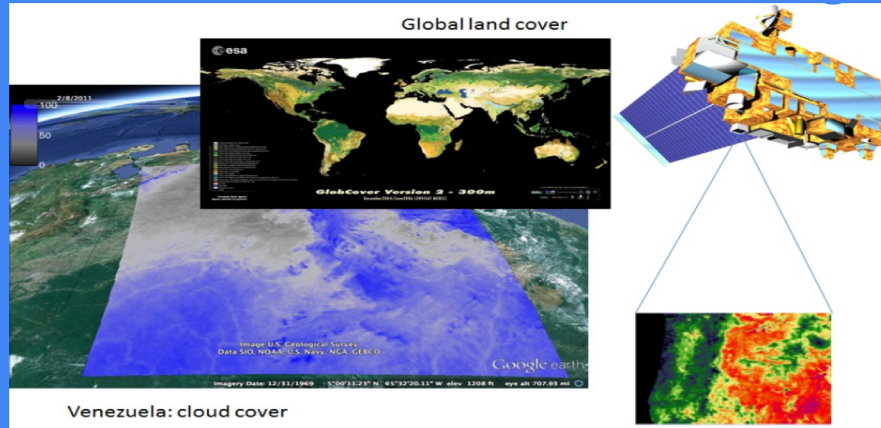
- User login:** A section on the left with fields for Username (jmalczyk) and Password, and buttons for Log in, Create new account, and Request new password.
- Search:** A central search bar containing "white stork" and a dropdown menu for "All Sensor Types". Below it is a "Search result" list with 16 entries, each with a checkbox and a magnifying glass icon. The entries include studies like "AbOVE: Peters Hebblewhite Alberta-BC Moose" and "HUJ MoveEcol Lab Israel: Foraging search efficiency".
- Map:** A large map on the right showing movement tracks (pink lines) across Europe and Africa. The tracks originate in Central Europe (Germany, Poland) and extend southwards through the Middle East and into East Africa (Ethiopia, Kenya, Tanzania). The map includes various geographical labels and a search bar at the top.

... and more movement data.

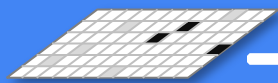




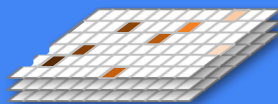
Combining biodiversity data with remote sensing products



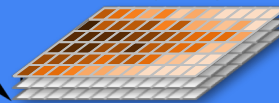
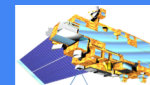
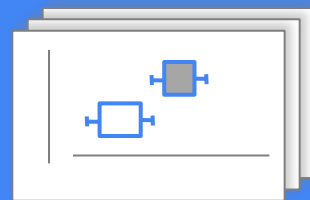
Species presence



Environment



Model



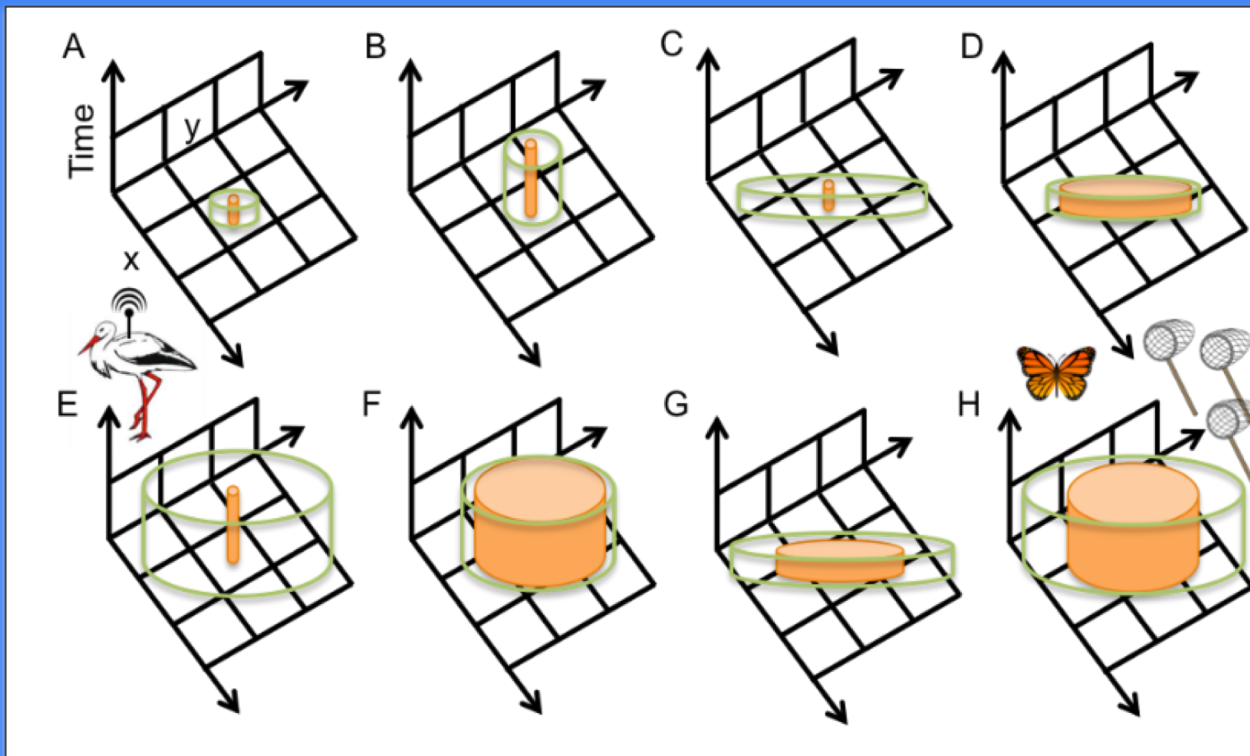
Prediction

Observation

Yale

Biodiversity data \cap Environmental data

Not trivial ...



Existing tools for data fusion



Desktop applications and libraries

Advantages

- Well documented
- Extensible
- Well integrated with other tools and systems
- Large community of developers and users
- What biology and ecology communities use and understand

Limits

- Requires everything (imagery and data) to be local
- Can't scale beyond local resources



ENV-Data

Advantages

- Large catalog of public imagery
- Kept up to date
- Specifically designed for annotating movement data

Limits

- Slow (hours - days)
- Not extensible
- No support for raster upload
- No support for spatial or temporal aggregation



Google Earth Engine

Advantages

- Large and growing catalog of imagery
- Kept up to date
- Abstracts complexity away (scales compute, manages tasking)
- Full spatial analysis API - supports aggregation in space-time

Limits

- High vendor lock-in / low portability
- Limited interoperability with traditional tools
- Very limited server-side logging (difficult to debug)
- Quota limited per user
- Fixed node size
- Mediocre performance across large vector datasets.



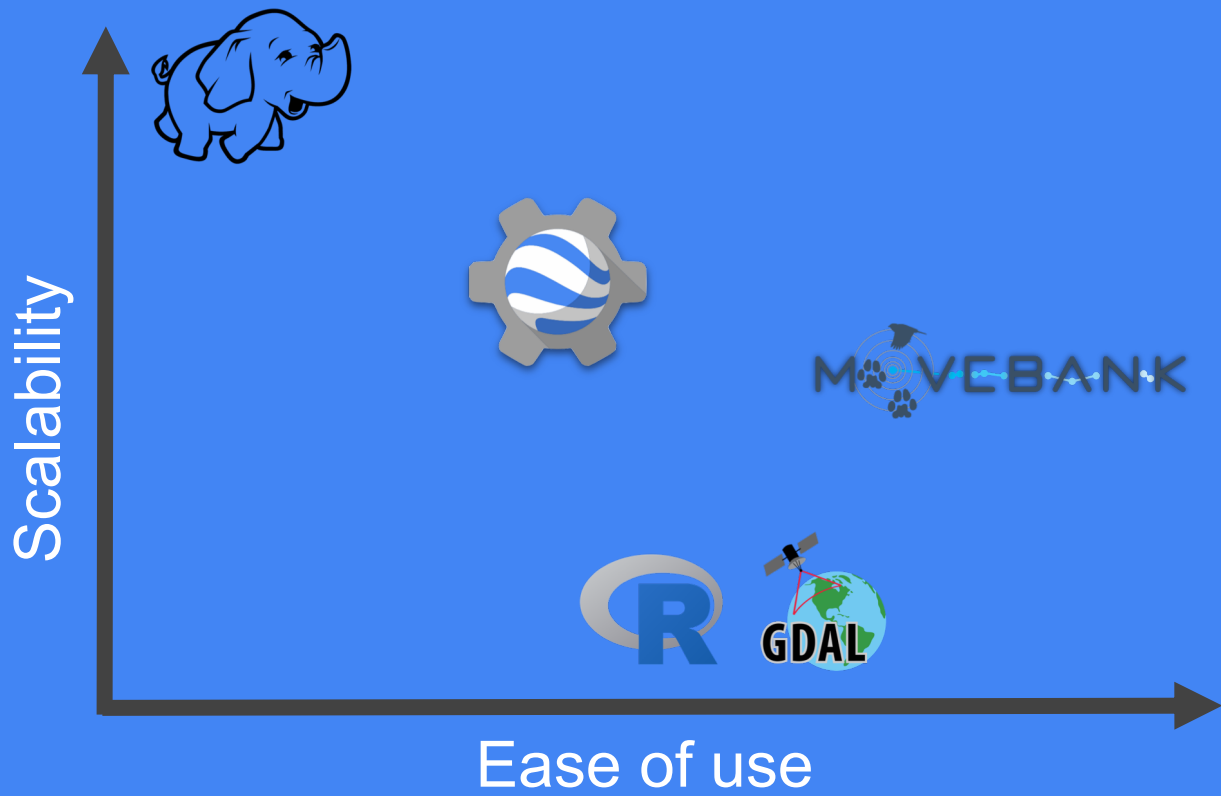
Big-data frameworks

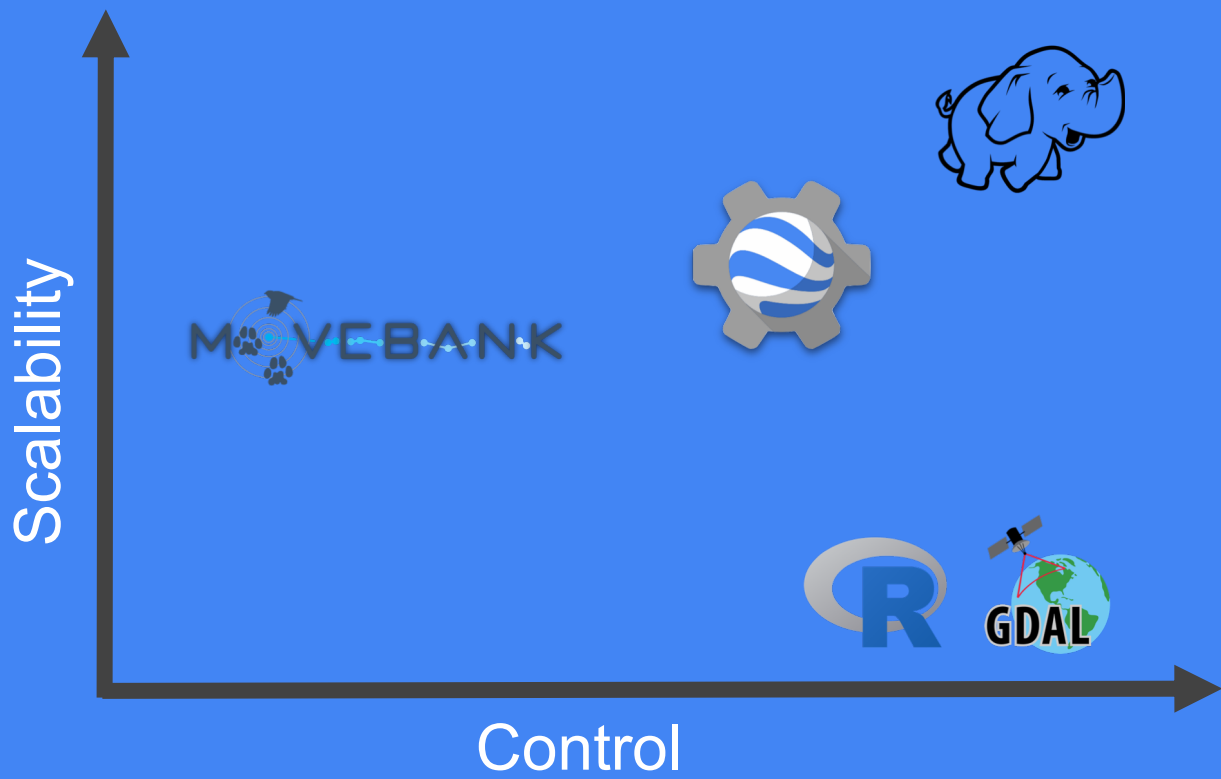
Advantages

- Ultimate flexibility
- Ultimate scalability

Limits

- Harder to use, less accessible to non-technical users
- Often limited documentation and support for spatial analysis





Bridging the divide with Apache Beam

Yale



beam

An advanced unified programming model

Implement batch and streaming data processing jobs that
run on any execution engine.



Yale

The Beam API

```
# sample images and apply spatiotemporal reducers
samples = (features
  | 'sample_pixels' >> beam.ParDo(sample_region, args, asset)
  | 'apply_reducers' >> beam.CombinePerKey(ReducePixels(args))
  | 'format_reducer_output' >> beam.ParDo(format_reducer_output)
  | 'group_by_location' >> beam.GroupByKey()
)
```

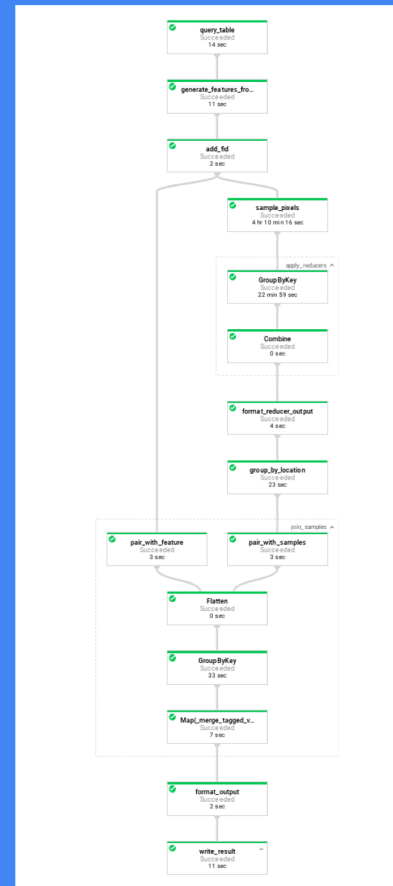


Cloud Dataflow features

Live pipeline monitoring

Detailed logging

Dataflow Step, 2018-06-12_19_35_04-12251...	worker	Any log level	No limit	Jump to now
Showing logs from all time (EDT) View				
▶		2018-06-13 01:13:08.739 EDT	Completed workitem: 3451789027802838680 in 2.474665165 seconds	
▶		2018-06-13 01:13:08.711 EDT	Finished processing workitem 3451789027802838680 successfully. Report	
▶		2018-06-13 01:13:08.558 EDT	Finished the size estimation of the input at 0 files. Estimation took	
▶		2018-06-13 01:13:08.506 EDT	Attempting refresh to obtain initial access_token	
▶		2018-06-13 01:13:08.506 EDT	Starting the size estimation of the input	
▶		2018-06-13 01:13:08.504 EDT	Renamed 15 shards in 1.35 seconds.	



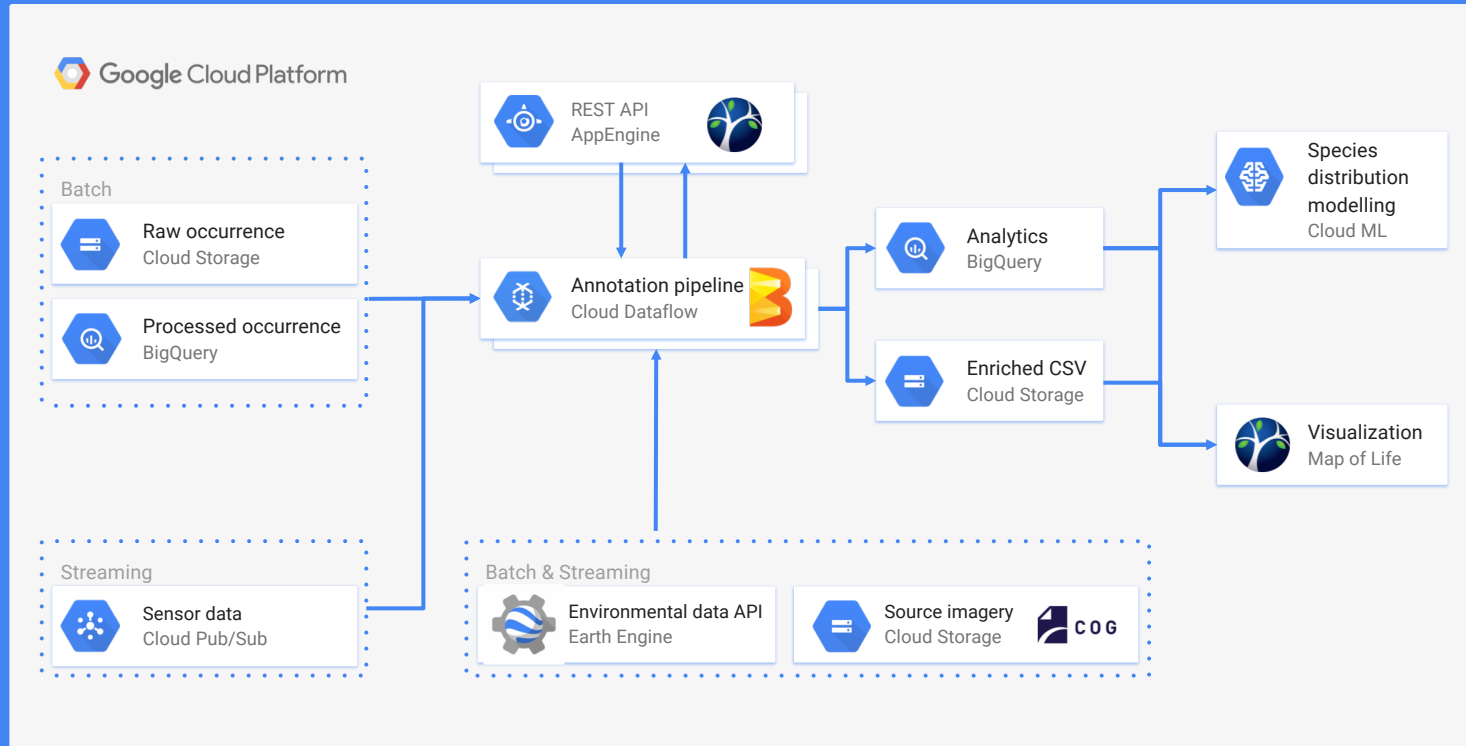
Autoscaling



Drawbacks

- Very new
- Need high availability access to pixel data (high QPS API or local to compute node)
- Cloud runners (DataFlow) have ~6 minute startup cost
- Less suitable for small requests
- Exotic environments difficult to support and scale
- Google Cloud Dataflow only full-service option

Architecture: Environmental annotation of biodiversity occurrence data



Project deliverables

- Open source Apache Beam code to run data fusion requests on local and API accessible datasets
- HTTP API to manage data fusion requests on the Google Cloud Dataflow pipeline runner
- Command-line interface to interact with API
- Web front-end
- Suite of 1km products suitable for conservation science (1km daily temperature and precip)