BEACON

ACCELERATING ACCESS TO MULTIDISCIPLINARY DATA WITH RELATIVE OPTIMIZED CHUNKING TECHNOLOGY

Tjerk Krijger (MARIS)

Funded by
the European Union
NextGenerationEU

EGU25 - 30 April 2025

ORIGIN OF BEACON

Challenge:

- Large collections of observation data (e.g. SeaDataNet CDI) can contain millions of files.
- Access to files (datasets) is possible, but data sub-setting proves to be difficult.
- How to optimize the systems for Machine2Machine access to subsets, enabling easy access to Jupyter Notebooks and other applications.
- How to go from files to serving applications as an actual “Data lake”?

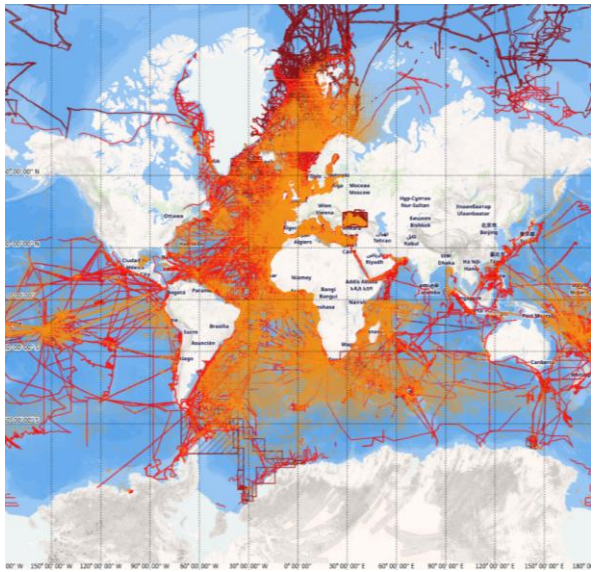
Example:

- Request: Give me all the temperature data in the North Sea, from 2000-2010, in degrees Celsius, at a depth of 0-5 m.
- Response: One NetCDF file containing exactly this data, which is then on the fly, directly usable in a Jupyter notebook and for HPC.

The goal of BEACON is to provide an easy-to-use, fast, reliable, and scalable solution for storing, processing, and retrieving large amounts of climate data.

Beacon has been part of various EU projects such as PHIDIAS HPC & EOSC-Future and is currently part of FAIR-EASE, Blue-Cloud 2026 and national initiatives.

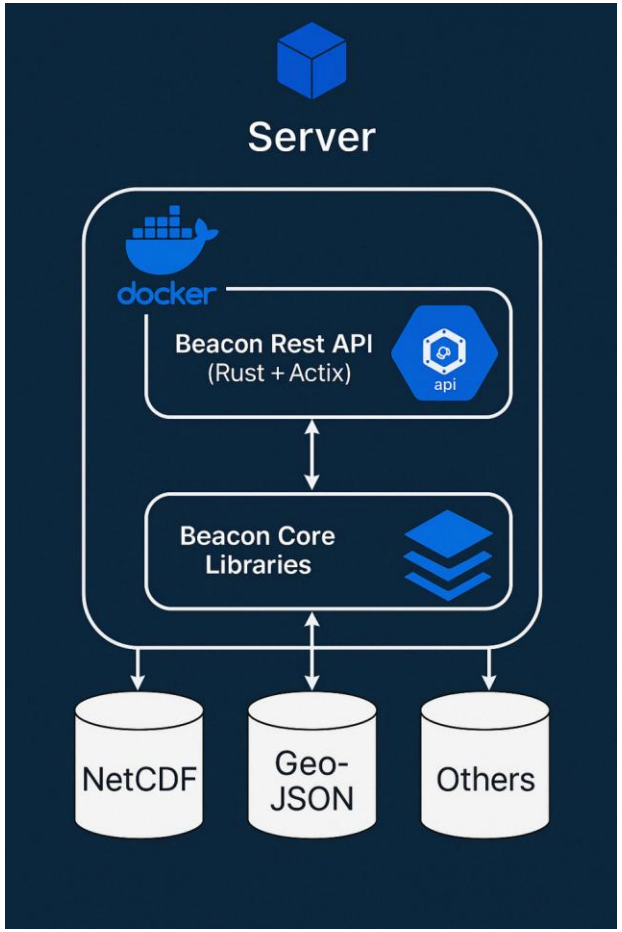
SeaDataNet CDI



Funded by the European Union
NextGenerationEU

BEACON IN A NUTSHELL

- Written in Rust + C
- High Performance Data Lake
- Runs on:
  - Linux
  - Windows
  - Docker Containers
- Consists of:
  - Rest API
  - Core Libraries
  - Real-Time sub-setting
  - Data harmonization (single output file)
  - Dynamic Chunking
- Produces different output formats:
  - NetCDF
  - CSV
  - JSON
  - IPC (Apache Arrow)
  - GeoJSON
  - Parquet
  - BBF (Beacon Binary Format)



- Handle any NetCDF Structure (E.g. Timeseries, Cruises, Gridded)
- Powerful query capabilities (SQL)
- Filter on:
  - Ranges
  - Polygons
  - Metadata
  - Union/Aggregation Queries
  - Federated Queries

Funded by
the European Union
NextGenerationEU

PERFORMANCE AND USAGE BEFORE CHUNKING APPLIED

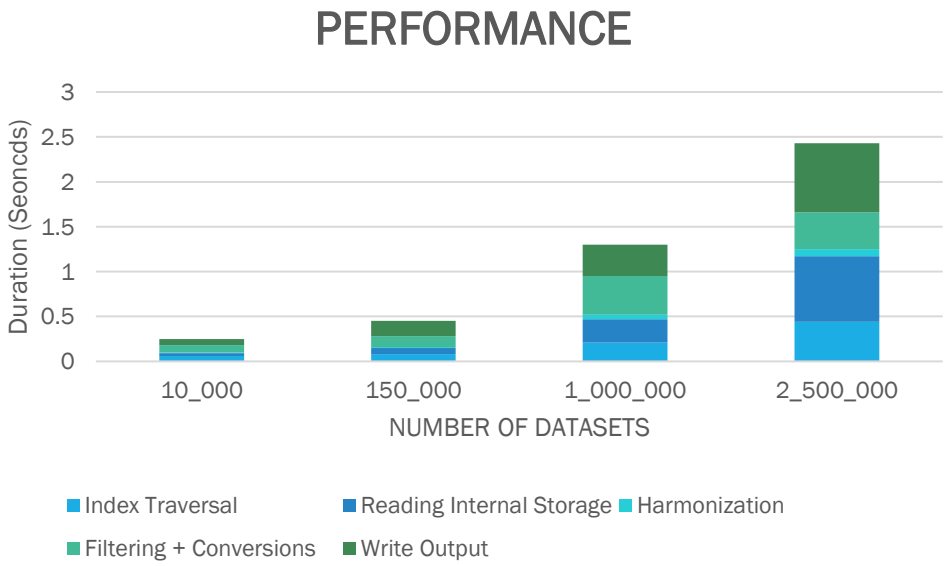
Performance

Loaded into BEACON all SDN CDI records:

- 2.5 millions datasets
- > 4 billion data points
- 200GB of NetCDF Data

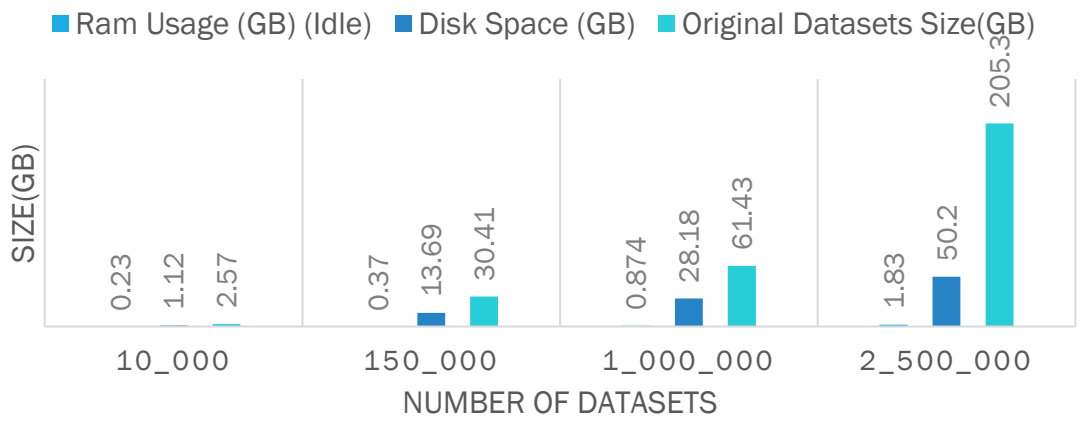
Query:

- Longitude from -8 to 12
- Latitude from 50 to 61
- Depth from 0 to 50
- Time from 2010 to 2012
- All the temperature parameters aggregated and harmonized in degrees Celsius
- Result: 12M points!



Beacon System Usage

- Can run on Laptops, Home PC's and Servers
- Only uses what's necessary to process the query



Beacon ERA5
11TB of Data
11K daily 0.05 gridded SST datasets.
260 Billion measurements
Resource Usage:
1.2GB of RAM,
250GB of Disk Space

Beacon Argo Data
600GB of NetCDF Data.
3.3M datasets.
Resource Usage:
600MB of RAM
50 GB of Disk Space

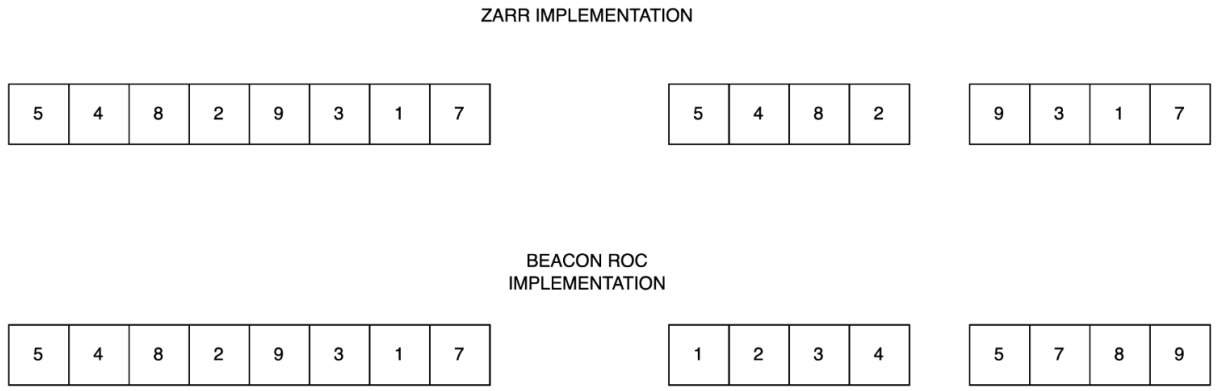
Beacon SeaDataNet
250 GB of NetCDF Data.
2.8M datasets.
Resource Usage:
1.8GB of RAM
60 GB of Disk Space

Funded by the European Union
NextGenerationEU

DYNAMIC CHUNKING

BEACON ROC (Relative Optimized Chunking):

- One BEACON instance can contain a collection of thousands of datasets
- Each dataset can be divided in X-number of chunks based on:
  - Columns (Time, Depth, etc.)
  - Geospatial chunking (Lon, Lat)



- ZARR but more powerful and faster
- Works for Gridded, Profiles, Cruises, etc.
- Data based approach instead of Dimension based approach

Depth	4	4	4	4	4	3	4	4
Longitude	10.5	11.2	8.7	11.2	11.4	6.5	12.1	12.1
Latitude	7.5	8.3	8.7	8.1	4.5	8.3	9.1	8.4

Basic Relative Chunk

Depth	3	4	4	4
Longitude	6.5	8.7	10.5	11.4
Latitude	8.3	8.7	7.5	4.5

Geo Relative Chunk

Chunk that has been chunked on how close data is relative to each other

Depth	4	4	4	4
Longitude	11.2	11.2	12.1	12.1
Latitude	8.1	8.3	8.4	9.1

Funded by the European Union
NextGenerationEU

CHUNKING ON TIME

Date & Time	SST (°C)
2024-08-17 14:22	22.4
2025-03-04 06:15	15.1
2024-11-09 23:50	17.8
2024-06-30 09:05	20.2
2025-01-18 04:40	13.7
2024-09-22 11:30	21.6
2025-02-11 01:20	14.5
2024-12-12 16:45	18.9
2024-07-05 20:10	19.5
2025-04-01 02:35	12.9
2024-10-03 07:55	16.3
2024-08-01 18:25	23.1
2025-03-22 13:10	16.8
2024-06-10 21:40	20.7
2024-09-15 10:00	22.0

One dataset converted into approximately equally sized chunks based on time

Chunk	Date & Time	SST (°C)
1	2024-06-10 21:40	20.7
	2024-06-30 09:05	20.2
	2024-07-05 20:10	19.5
2	2024-08-01 18:25	23.1
	2024-08-17 14:22	22.4
	2024-09-15 10:00	22.0
3	2024-09-22 11:30	21.6
	2024-10-03 07:55	16.3
	2024-11-09 23:50	17.8
4	2024-12-12 16:45	18.9
	2025-01-18 04:40	13.7
	2025-02-11 01:20	14.5
5	2025-03-04 06:15	15.1
	2025-03-22 13:10	16.8
	2025-04-01 02:35	12.9

Funded by the European Union
NextGenerationEU

CHUNKING ON TIME FIRST, DEPTH SECOND

Date	Depth (m)	SST (°C)
2024-08-03	10	19.8
2024-08-03	50	14.6
2024-06-21	25	17.9
2024-07-10	80	11.2
2024-09-14	5	22.1
2024-10-01	60	13.4
2024-10-01	90	10.3
2024-06-21	70	12.7
2024-07-10	35	15.5
2024-09-14	15	20.0
2024-12-05	40	14.0
2024-12-20	100	9.3
2024-11-11	30	16.1
2024-12-05	20	18.2
2024-08-17	45	15.0

One dataset converted into approximately equally sized chunks based on firstly time and secondly depth

Chunk	Date	Depth (m)	SST (°C)
1	2024-06-21	25	17.9
	2024-06-21	70	12.7
	2024-07-10	35	15.5
2	2024-07-10	80	11.2
	2024-08-03	10	19.8
	2024-08-03	50	14.6
3	2024-08-17	45	15.0
	2024-09-14	5	22.1
	2024-09-14	15	20.0
4	2024-10-01	60	13.4
	2024-10-01	90	10.3
	2024-11-11	30	16.1
5	2024-12-05	20	18.2
	2024-12-05	40	14.0
	2024-12-20	100	9.3

Funded by the European Union NextGenerationEU

GEOSPATIAL CHUNKING

Date & Time	Depth (m)	SST (°C)	Longitude	Latitude
2024-07-05 06:45	5	21.3	-73.2041	-12.3847
2024-08-11 14:30	30	16.2	149.3872	-3.4501
2024-09-19 03:15	20	18.7	-25.1739	34.2185
2024-10-07 22:00	50	14.5	43.1874	-8.6032
2024-11-02 10:20	10	19.9	-122.9154	36.2248
2024-12-18 09:40	80	12.1	88.5302	20.1487
2025-01-23 18:10	100	9.6	-3.1947	60.3924
2025-02-05 12:35	15	20.8	110.2381	17.4593
2025-03-10 07:50	60	13.8	-54.9875	-25.1012
2025-04-04 00:25	25	17.5	13.8765	42.0500
2025-04-04 00:45	25	17.3	13.8790	42.0523
2024-08-11 08:10	5	22.4	149.3895	-3.4479
2024-09-19 05:30	100	10.1	-25.1783	34.2198
2024-12-18 15:30	70	13.0	88.5284	20.1520
2025-01-23 05:00	20	18.2	-3.1971	60.3901

One dataset converted into geospatial chunks

Chunk	Date & Time	Depth (m)	SST (°C)	Longitude	Latitude
1	2024-07-05 06:45	5	21.3	-73.2041	-12.3847
	2024-11-02 10:20	10	19.9	-122.9154	36.2248
	2025-03-10 07:50	60	13.8	-54.9875	-25.1012
2	2024-10-07 22:00	50	14.5	43.1874	-8.6032
	2024-12-18 09:40	80	12.1	88.5302	20.1487
	2025-02-05 12:35	15	20.8	110.2381	17.4593
	2024-12-18 15:30	70	13.0	88.5284	20.1520
3	2024-09-19 03:15	20	18.7	-25.1739	34.2185
	2024-09-19 05:30	100	10.1	-25.1783	34.2198
4	2024-08-11 14:30	30	16.2	149.3872	-3.4501
	2024-08-11 08:10	5	22.4	149.3895	-3.4479
5	2025-01-23 18:10	100	9.6	-3.1947	60.3924
	2025-04-04 00:25	25	17.5	13.8765	42.0500
	2025-04-04 00:45	25	17.3	13.8790	42.0523
	2025-01-23 05:00	20	18.2	-3.1971	60.3901

Funded by the European Union NextGenerationEU

GEOSPATIAL CHUNKING

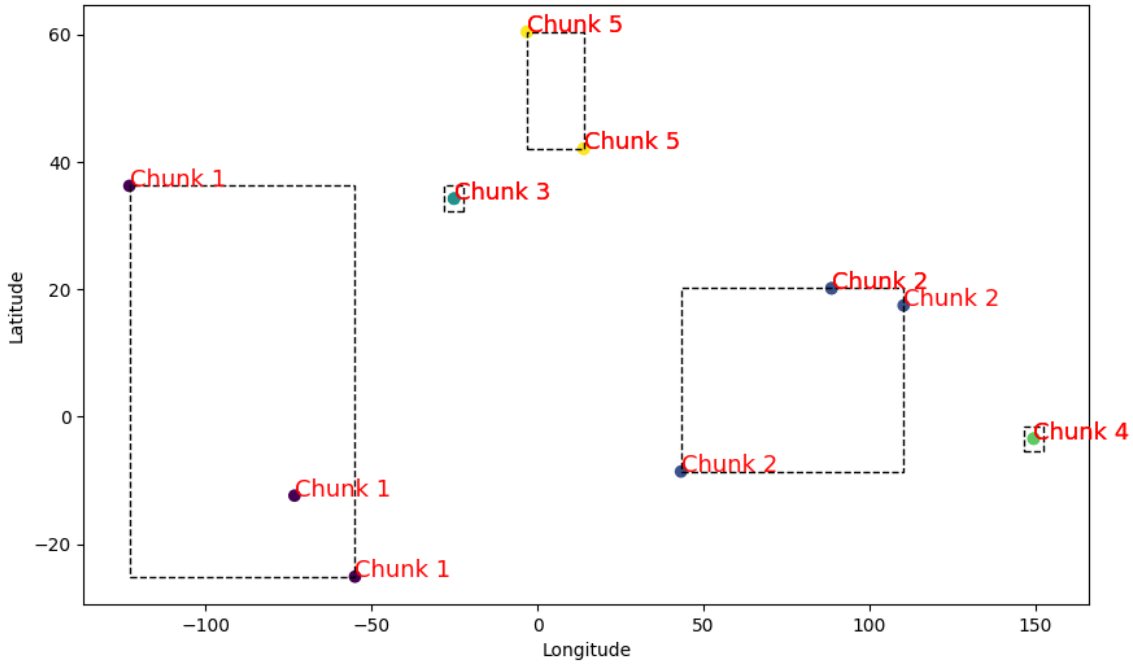
Based on R-tree indexing

- Balanced tree used for indexing multidimensional information
- Grouping of objects and represented by their minimum bounding rectangles (MBRs)
- Builds hierarchy based on proximity

Application in BEACON

- One BEACON instance can contain a collection of datasets
- Each dataset can be divided in X-number of chunks
- X is a balanced number depending on the dataset size
  - E.g. dataset with 256 rows is divided in X = 8 chunks of approx. 32 rows each

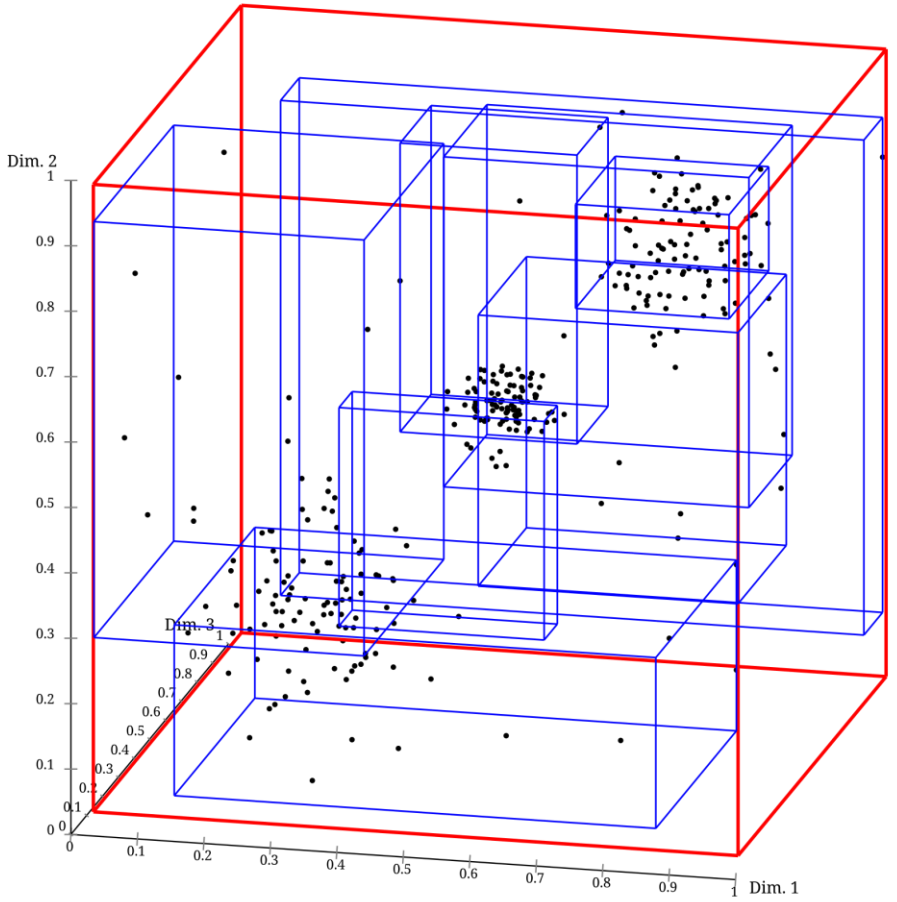
Previous example visualized



Funded by the European Union
NextGenerationEU

GEOSPATIAL CHUNKING IN 3D

- Geospatial chunks can also be combined with a third dimension
- Order of chunking can be determined
  - E.g. first chunk on depth, then geospatially, or other way around
  - Not limited to time, depth, geospatial, can also be applied to all other columns in the datasets such as platform type.

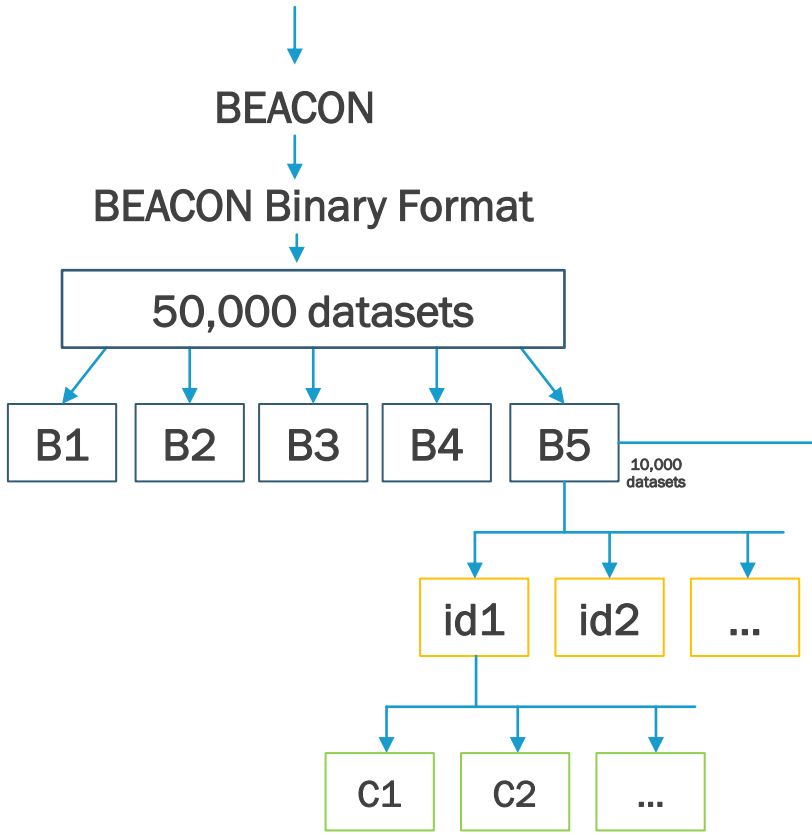


Funded by the European Union
NextGenerationEU

HOW DOES IT WORK IN PRACTICE?

Query to BEACON example instance:

Time range [2000-2010], depth range [0-5m], Temperature, BBOX [52.1, 52.3, 3.9, 4.2]



Chunking speeds up the query by reducing I/O

Up to 90% faster (depending on the datasets & query)

Index

Lon	Lat	Time	...	Dataset id	Chunk
3.9	52.1	2004-01-02	...	id1	1
4.0	54.0	2004-05-05	...		2
5.1	55.3	2008-01-02	...	id2	1
..	..	..	...	..	..

BEACON example instance contains the following data collection:

- Temperature & Salinity data
- North Sea region
- 50,000 datasets
- 2 million measurements
- 2000-2025

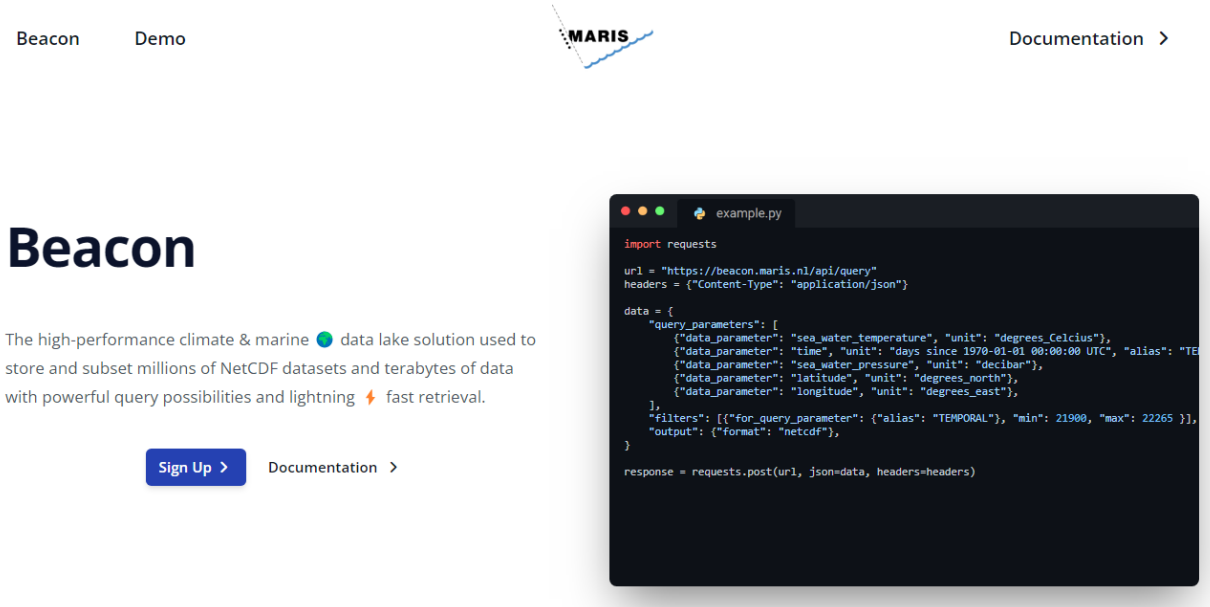
Manager configuration:

- Geospatial chunking
- Block size = 10,000 datasets

Funded by the European Union
NextGenerationEU

STATUS

- BEACON freely available at:
  - https://beacon.maris.nl/
- Docs: https://maris-development.github.io/beacon/
- GitHub: https://github.com/maris-development/beacon
- Allows you to set-up your own BEACON instance
  - With your data collection
- On BC2026 VRE multiple BEACON instances are set-up
  - With example notebooks



Funded by the European Union
NextGenerationEU

BEACON INSTANCES/NODES

- Euro-Argo data
  - *retrieved from S3 bucket*
- CORA Profile data
  - *retrieved from CMEMS, product: INSITU_GLO_PHY_TS_DISCRETE_MY_013_001*
- CORA Timeseries data
  - *retrieved from CMEMS, product: INSITU_GLO_PHY_TS_DISCRETE_MY_013_001*
- EMODnet Chemistry data
  - *retrieved from EMODnet Chemistry WebODV (Eutrophication_Atlantic_profiles_2023_unrestricted.odv)*
- WOD data
  - *retrieved from ncei.noaa.gov*
- SeaDataNet CDI
- CMEMS BGC data
  - *retrieved from CMEMS, product: INSITU_GLO_BGC_DISCRETE_MY_013_046*

Funded by the European Union NextGenerationEU

QUERYING BEACON

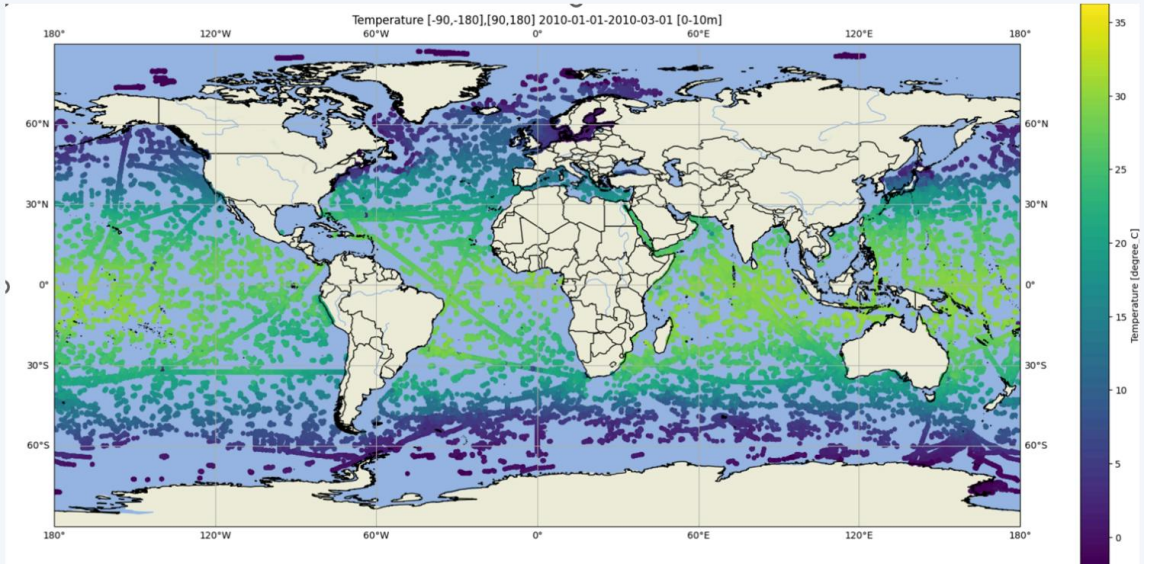
- Retrieve specific data based on your requirements
- JSON query object and sending a POST request to the query endpoint
  - Select Columns
  - Define filters for the columns
  - Select the output format
  - Query!
- Retrieve data that matches your specified parameters, units, filters, and other criteria

```
response = requests.post("https://beacon-wod.maris.nl/api/query", json.dumps(query_body), headers={
    "Authorization": f"Bearer {Token}",
    "Content-type": "application/json"
})
```

datetime	Temperature	Unit	DEPTH	LONGITUDE	LATITUDE	DATASET	cruise-identifier	cast	WMO_ID	country	dataset_id
2010-01-01	24.700001	degree_C	1.0	-95.000000	-5.000000	moored buoy	US015955	12327900	2.115095e+09	UNITED STATES	11382402
2010-01-01	26.120001	degree_C	5.0	-110.000000	0.000000	moored buoy	US015969	15272646	2.115095e+09	UNITED STATES	11382406
2010-01-01	26.110001	degree_C	10.0	-110.000000	0.000000	moored buoy	US015969	15272646	2.115095e+09	UNITED STATES	11382406
2010-01-01	28.959999	degree_C	1.5	147.039993	4.960000	moored buoy	JP020811	12327885	2.115114e+09	JAPAN	11382401
2010-01-01	25.440001	degree_C	1.0	-37.880001	20.040001	moored buoy	US026451	12327924	2.115104e+09	UNITED STATES	11382426
...	...	...	...	...	...	...	...	...	...	...	...
2010-03-01	29.379999	degree_C	1.0	-34.639999	-18.860001	moored buoy	BR001204	12332326	2.115093e+09	BRAZIL	11390007
2010-03-01	27.370001	degree_C	1.0	-155.009995	7.980000	moored buoy	US015934	12332305	2.115114e+09	UNITED STATES	11389980
2010-03-01	29.860001	degree_C	1.0	-170.020004	-2.170000	moored buoy	US015927	12332281	2.115114e+09	UNITED STATES	11389956
2010-03-01	29.540001	degree_C	1.0	89.989998	-1.670000	moored buoy	JP032626	12327389	2.115115e+09	JAPAN	11389989
2010-03-01	28.430000	degree_C	1.0	165.130005	8.050000	moored buoy	US015890	12332307	2.115114e+09	UNITED STATES	11389966

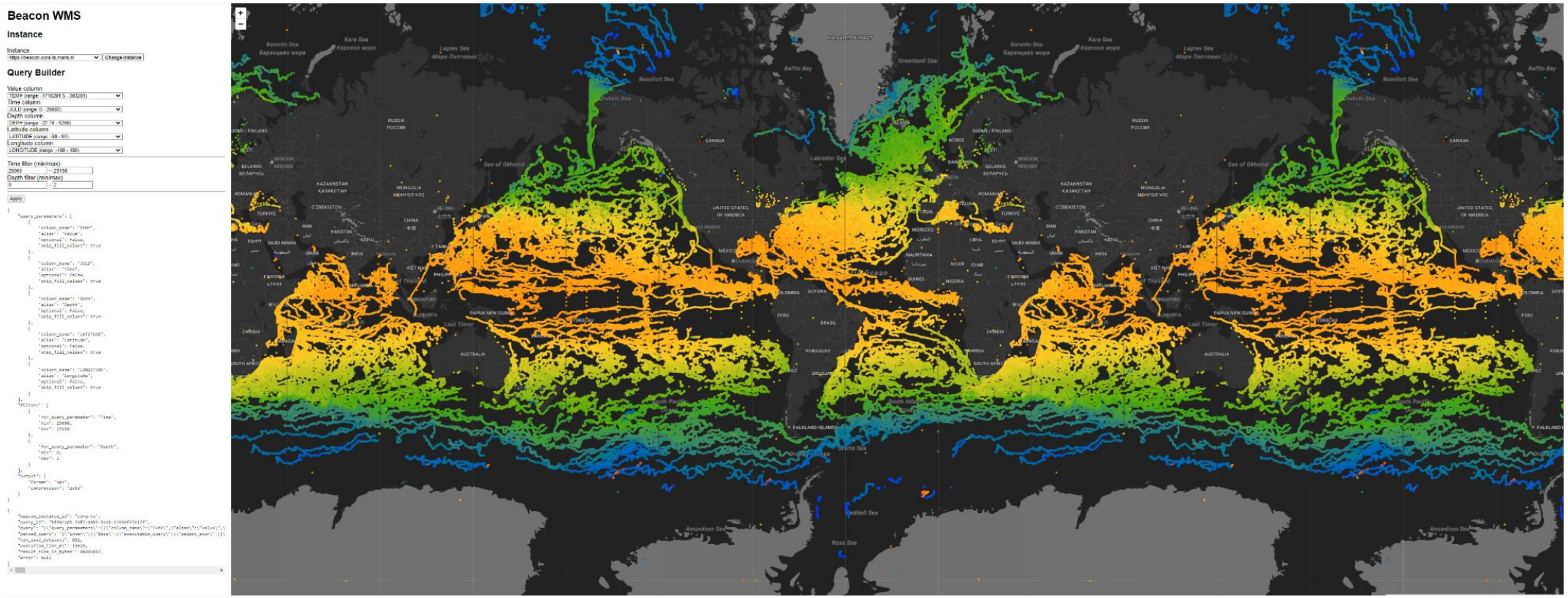
308863 rows × 11 columns

In this example we can then very quickly plot this subset of the temperature data between 0-10 meters depth from the WOD collection of the first months of 2010 produced.



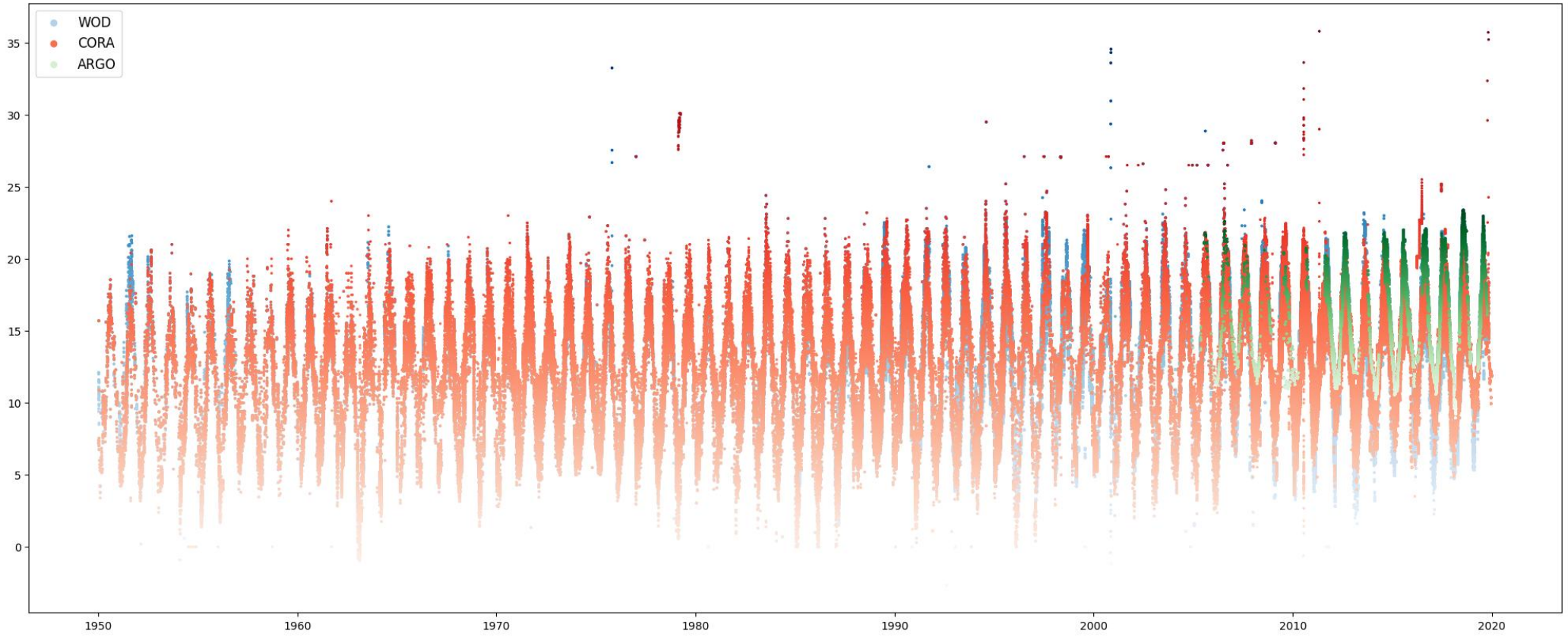
Funded by the European Union
NextGenerationEU

EXAMPLE: (DRAWING 20M TEMPERATURE POINTS)



Funded by the European Union
NextGenerationEU

EXAMPLE: 1950 - 2020 NORTH SEA TEMPERATURE



Funded by the European Union
NextGenerationEU

GET IN TOUCH

- Tjerk Krijger (tjerk@maris.nl)
- Robin Kooyman (robin@maris.nl)
- Dick Schaap (dick@maris.nl)
- Peter Thijsse (peter@maris.nl)

Funded by the European Union
NextGenerationEU