

Massively Scalable, Cloud-native Hydrographic Data Processing

From Sensor Data to High-Resolution
Deliverables

HYDRO 2025

29th of October 2025

01

Introduction

Information about the company and background

The Ocean Big Data Specialists

Building Technology
for managing,
visualizing and
processing ocean big
data at scale



Founded in 2011



Cloud & Big Data



50+ employees



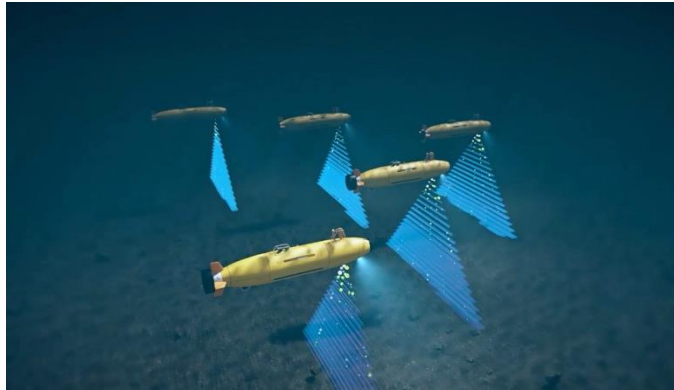
Located in Kiel, Germany



GLOBAL TOP 100 GEOSPATIAL COMPANIES IN 2025

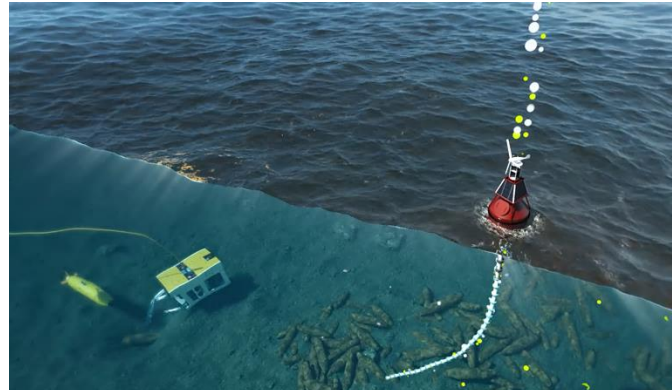
Megatrends in Ocean Data

Ocean data handling needs new paradigms



Autonomy

Autonomous systems and next-gen sensors are redefining data collection at sea



Connectivity

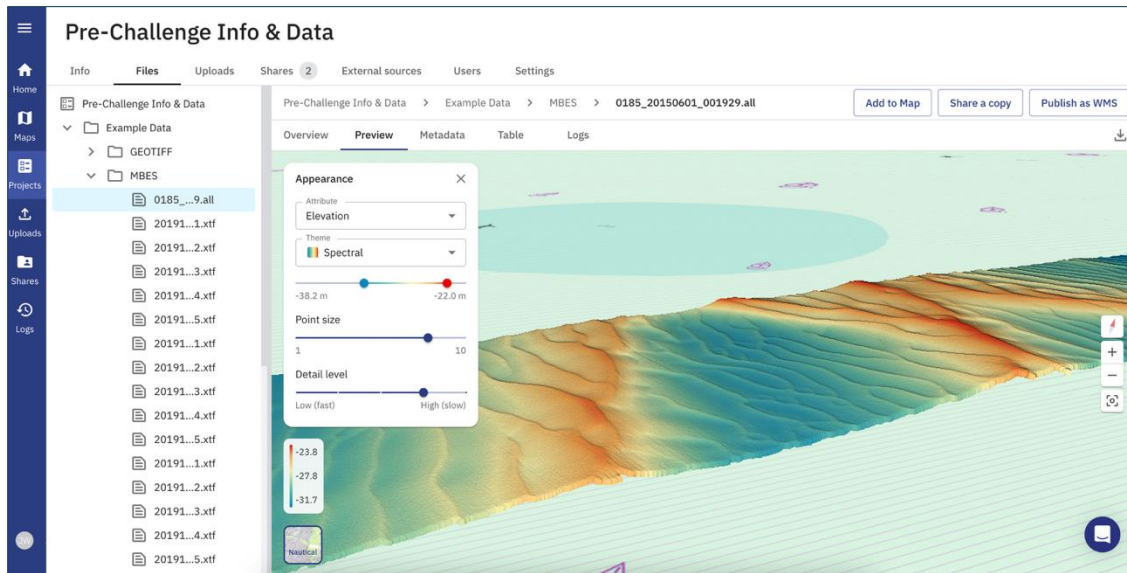
Connectivity and bandwidth is increasing, enabling smarter data workflows



Big Data & AI

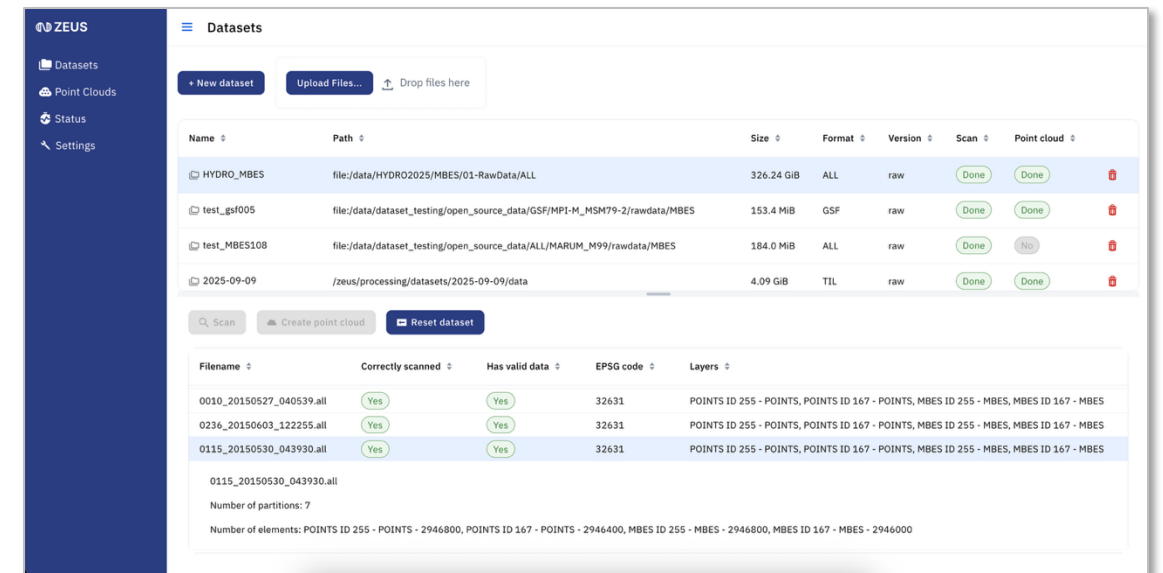
Ocean data is growing fast in volume, variety, and velocity — unlocking potential for Artificial Intelligence

Our Technology



Ocean Data Platform

Managing, sharing and visualizing large volumes of hydrographic and geospatial data



The screenshot displays the 'Datasets' interface. It includes a sidebar with navigation options like 'Datasets', 'Point Clouds', 'Status', and 'Settings'. The main area shows a table of datasets with columns for Name, Path, Size, Format, Version, Scan, and Point cloud. Below the table, there are buttons for 'Scan', 'Create point cloud', and 'Reset dataset'. A detailed view of a dataset is shown below the table, including its filename, processing status, EPSG code, and layers.

Name	Path	Size	Format	Version	Scan	Point cloud
HYDRO_MBES	file:/data/HYDRO2025/MBES/01-RawData/ALL	326.24 GiB	ALL	raw	Done	Done
test_gs4005	file:/data/dataset_testing/open_source_data/GSF/MPI-M_MSM79-2/rawdata/MBES	153.4 MiB	GSF	raw	Done	Done
test_MBES108	file:/data/dataset_testing/open_source_data/ALL/MARUM_M99/rawdata/MBES	184.0 MiB	ALL	raw	Done	No
2025-09-09	/zeus/processing/datasets/2025-09-09/data	4.09 GiB	TIL	raw	Done	Done

Filename	Correctly scanned	Has valid data	EPSG code	Layers
0010_20150527_040539.all	Yes	Yes	32631	POINTS ID 255 - POINTS, POINTS ID 167 - POINTS, MBES ID 255 - MBES, MBES ID 167 - MBES
0236_20150603_122255.all	Yes	Yes	32631	POINTS ID 255 - POINTS, POINTS ID 167 - POINTS, MBES ID 255 - MBES, MBES ID 167 - MBES
0115_20150530_043930.all	Yes	Yes	32631	POINTS ID 255 - POINTS, POINTS ID 167 - POINTS, MBES ID 255 - MBES, MBES ID 167 - MBES

0115_20150530_043930.all
Number of partitions: 7
Number of elements: POINTS ID 255 - POINTS - 2946800, POINTS ID 167 - POINTS - 2946400, MBES ID 255 - MBES - 2946800, MBES ID 167 - MBES - 2946000

Ocean Data Processing Engine

Processing very large volumes of hydrographic and geospatial data in distributed environments

02

Why?

Do we need some new paradigms when dealing with ocean data?

Why

The need of something new?

- Being able to process increasing data sizes without limitations - Terabytes, Petabytes and even more
- Process and analyse data without loss of information
- Unify data from non-standardized, diverse binary data formats
- Open-source technology to allow compatibility with any other system
- Automate processing workflows to support people's work



02

Technical Background

Which technologies are used and why did we specifically decide for these

Key Factors

Technological choices have been made based on the factors scalability, independence and holistic

Scalable

Independent

Holistic

Scalability

We want to scale our computing and only be limited by the available hardware resources

Independence

We want to be able to run our technology everywhere at every time, and be interoperable with any other system

Holistic

Processing massive datasets as a whole is key for unlocking the full potential of advanced data analysis and AI approaches

Key Technologies



Resource management system for maximum efficiency of hardware usage



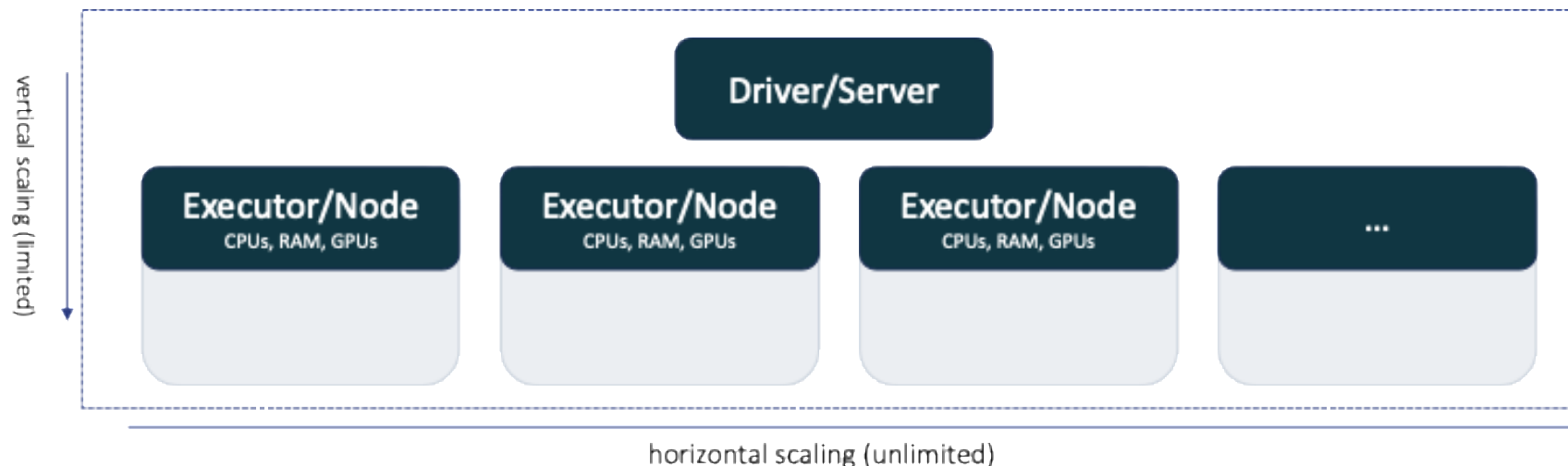
Distributed computing framework based on MapReduce paradigm



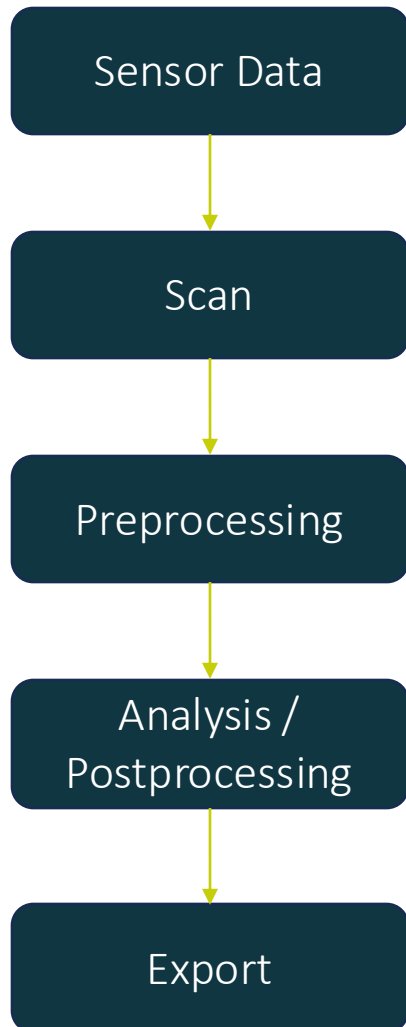
Open source data format highly optimized for big data operations

Cluster

A cluster consists of an unlimited amount of individual executors nodes



Processing Flow



Supported data formats for reading

- MBES: *.all, *.kmall, *.s7k, *.gsf, *.xtf, *.til, *.las/laz, *.xyz ; SSS: *.xtf, *.jsf, *.sdf, *.sds, *.til
- Raster: geotiff/cog, asciigrid, csv, geoparquet; Pointcloud: las/laz, pts, xyz, csv, geoparquet

Unification of marine binary sensor data formats

- using Kaitai binary parser for reading binary files
- compatible with almost all programming languages

Partitioning of data for parallel computing

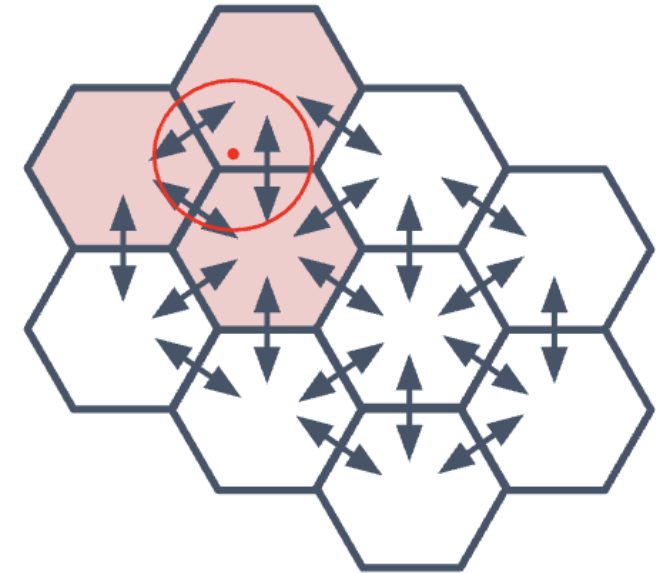
- based on the MapReduce programming model
- sensor specific preprocessing steps

Pointcloud based computations

- keeping all information for processing
- sensor independent “calculation library”

Supported data formats for writing

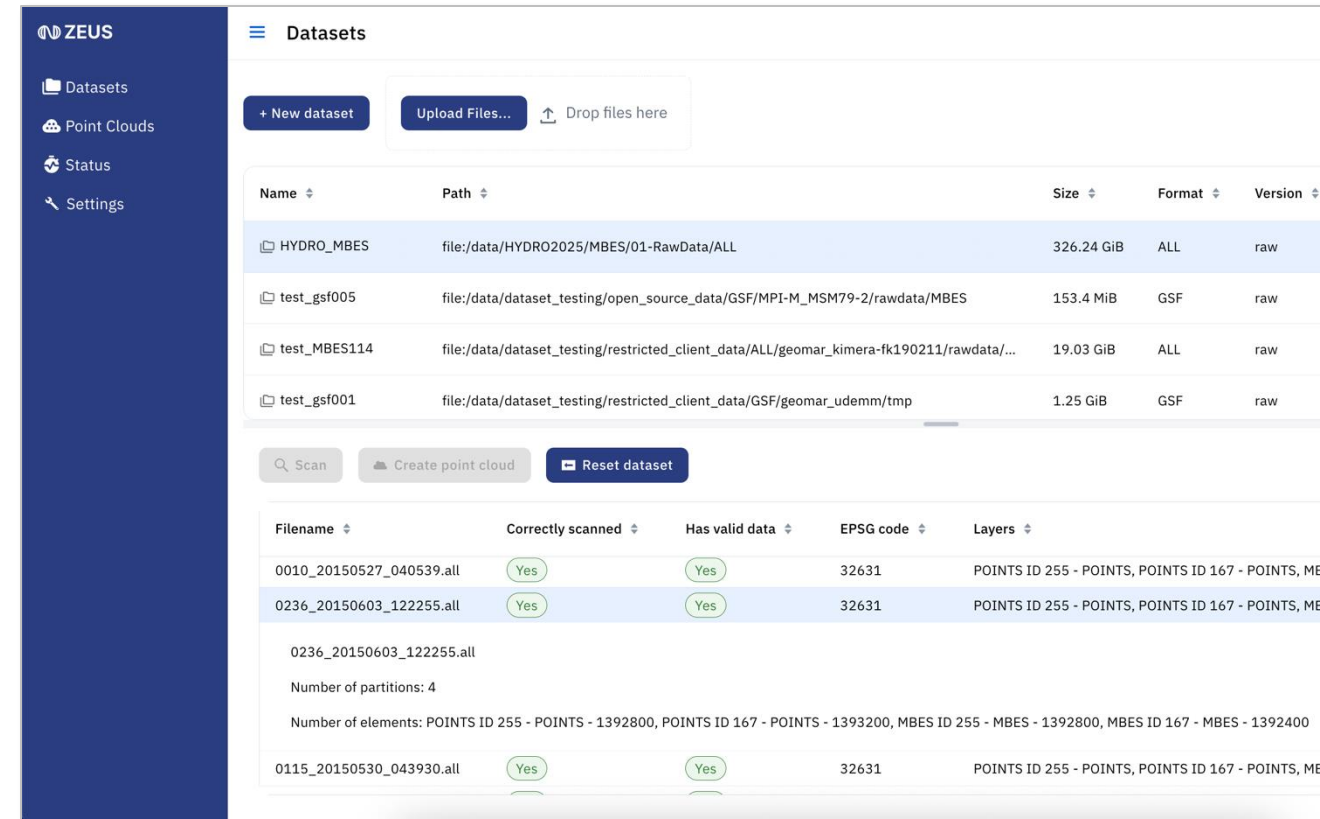
- Raster: geotiff/cog, asciigrid, csv, parquet/geoparquet, png, tms;
- Pointcloud: cesium, csv, parquet/geoparquet, las/laz



Conducting quality checks

- Is sensor data compliant with format specifications
→ Correctly scanned
- Are required datagrams for processing available
→ Has valid data
- Is CRS information available
→ EPSG Code
- Which data layers are contained (e.g., several frequencies in SSS data)
→ Layers

Data partitioning for parallel processing



The screenshot displays the ZEUS Datasets interface. On the left is a sidebar with navigation links: Datasets, Point Clouds, Status, and Settings. The main area is titled 'Datasets' and includes buttons for '+ New dataset', 'Upload Files...', and 'Drop files here'. Below this is a table listing datasets with columns for Name, Path, Size, Format, and Version.

Name	Path	Size	Format	Version
HYDRO_MBES	file:/data/HYDRO2025/MBES/01-RawData/ALL	326.24 GiB	ALL	raw
test_gsf005	file:/data/dataset_testing/open_source_data/GSF/MPI-M_MSM79-2/rawdata/MBES	153.4 MiB	GSF	raw
test_MBES114	file:/data/dataset_testing/restricted_client_data/ALL/geomar_kimera-fk190211/rawdata/...	19.03 GiB	ALL	raw
test_gsf001	file:/data/dataset_testing/restricted_client_data/GSF/geomar_udemm/tmp	1.25 GiB	GSF	raw

Below the table are buttons for 'Scan', 'Create point cloud', and 'Reset dataset'. A second table provides a detailed view of a dataset, showing columns for Filename, Correctly scanned, Has valid data, EPSG code, and Layers.

Filename	Correctly scanned	Has valid data	EPSG code	Layers
0010_20150527_040539.all	Yes	Yes	32631	POINTS ID 255 - POINTS, POINTS ID 167 - POINTS, ME
0236_20150603_122255.all	Yes	Yes	32631	POINTS ID 255 - POINTS, POINTS ID 167 - POINTS, ME
0236_20150603_122255.all				
Number of partitions: 4				
Number of elements: POINTS ID 255 - POINTS - 1392800, POINTS ID 167 - POINTS - 1393200, MBES ID 255 - MBES - 1392800, MBES ID 167 - MBES - 1392400				
0115_20150530_043930.all	Yes	Yes	32631	POINTS ID 255 - POINTS, POINTS ID 167 - POINTS, ME

Unified reader system for diverse set of marine sensor data

Preprocess

Raster/Pointcloud

Data Type	Data Sources
Ping Data	<i>Datagram XYZ</i>
Coordinates	<i>Datagram Position</i>
Attitude	<i>Datagram Attitude</i>

Preprocessing Step
Conversion to unified parquet

SSS

Data Type	Data Sources
Ping Data	<i>Datagram XYZ</i>
Coordinates	<i>Datagram Position</i>
Attitude	<i>Datagram Attitude</i>

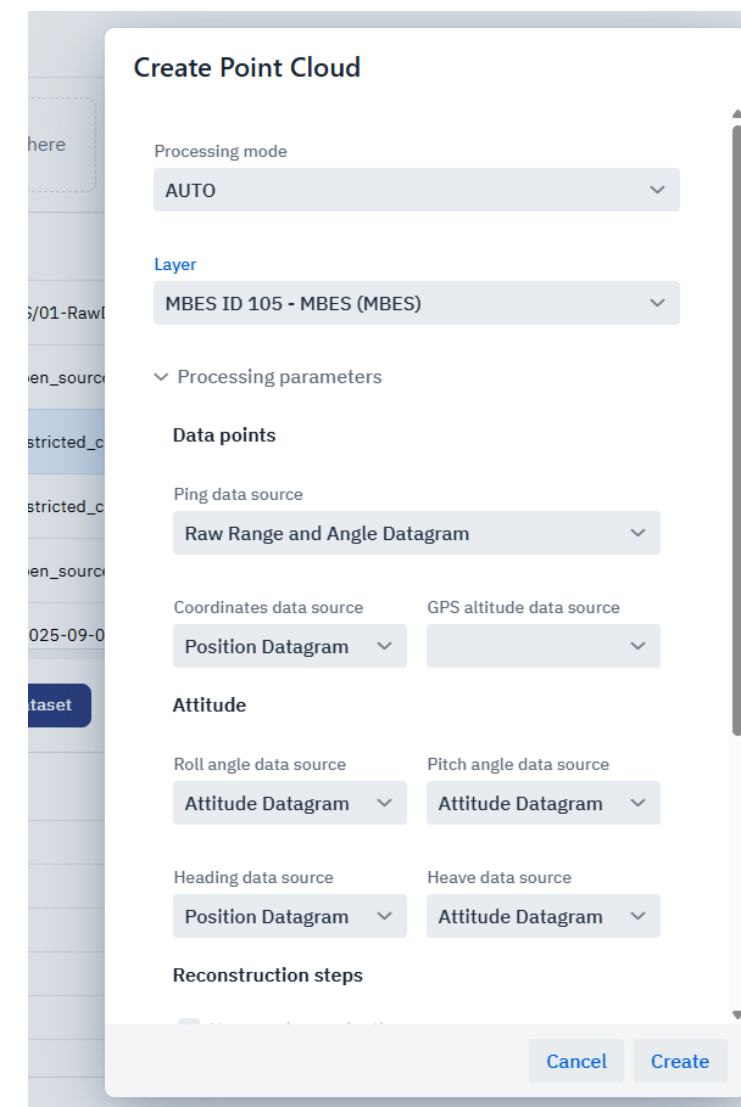
Preprocessing Step
Bottom detection
Slant range correction
Ground range correction
Interpolate GPS & smooth heading
Georeferencing
Conversion to unified parquet

MBES

Data Type	Data Sources
Ping Data	<i>Datagram XYZ</i>
Coordinates	<i>Datagram Position</i>
Altitude	<i>Datagram Position</i>
Attitude	<i>Datagram Attitude</i>

Preprocessing Step
Interpolate GPS
Correct for SVPs
Correct sensor offsets
Correct rotation
Correct heave
Conversion to unified parquet

Two processing modes: „AUTO“ and „MANUAL“ with sensor specific processing steps



The screenshot shows a 'Create Point Cloud' dialog box with the following settings:

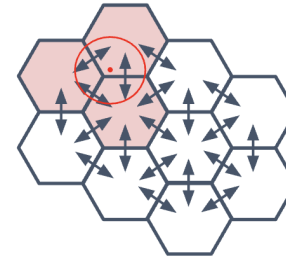
- Processing mode:** AUTO
- Layer:** MBES ID 105 - MBES (MBES)
- Processing parameters:**
 - Data points:**
 - Ping data source:** Raw Range and Angle Datagram
 - Coordinates data source:** Position Datagram
 - GPS altitude data source:** (empty)
 - Attitude:**
 - Roll angle data source:** Attitude Datagram
 - Pitch angle data source:** Attitude Datagram
 - Heading data source:** Position Datagram
 - Heave data source:** Attitude Datagram
 - Reconstruction steps:** (empty)

Buttons at the bottom: Cancel, Create

Flexible “calculation library” independent of input sensor due to unified data structure

Calculations

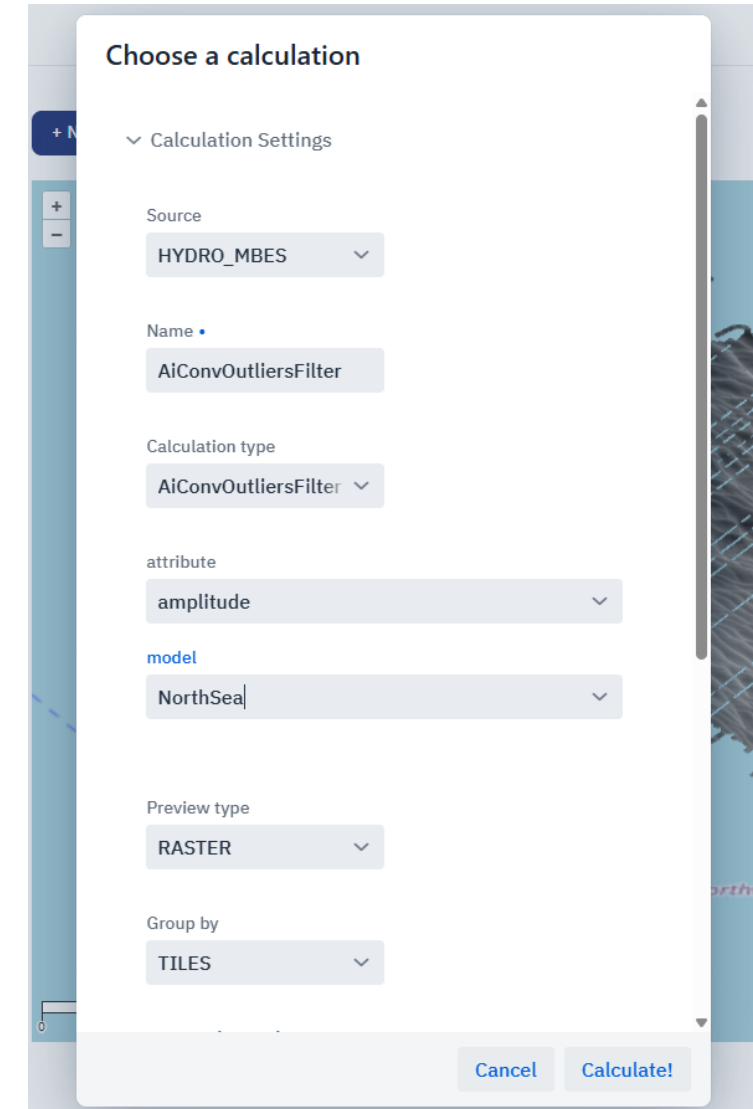
- data quality checks (point spacing, footprints, coverage, ...)
- filtering (statistical, AI-based)
- neighbourhood analysis (gradient, tpi, curvature, ...)
- anomaly detection (statistical, AI-based)



Interoperable with Python libraries for advanced analytics

Custom developed algorithm for nearest neighbour search

- up to 9× faster than existing methods (KNN in Spark MLlib, R² in Apache Sedona, FAISS, etc.)



Choose a calculation

Calculation Settings

Source: HYDRO_MBES

Name: AIConvOutliersFilter

Calculation type: AIConvOutliersFilter

attribute: amplitude

model: NorthSea

Preview type: RASTER

Group by: TILES

Cancel Calculate!

03

Benchmarking

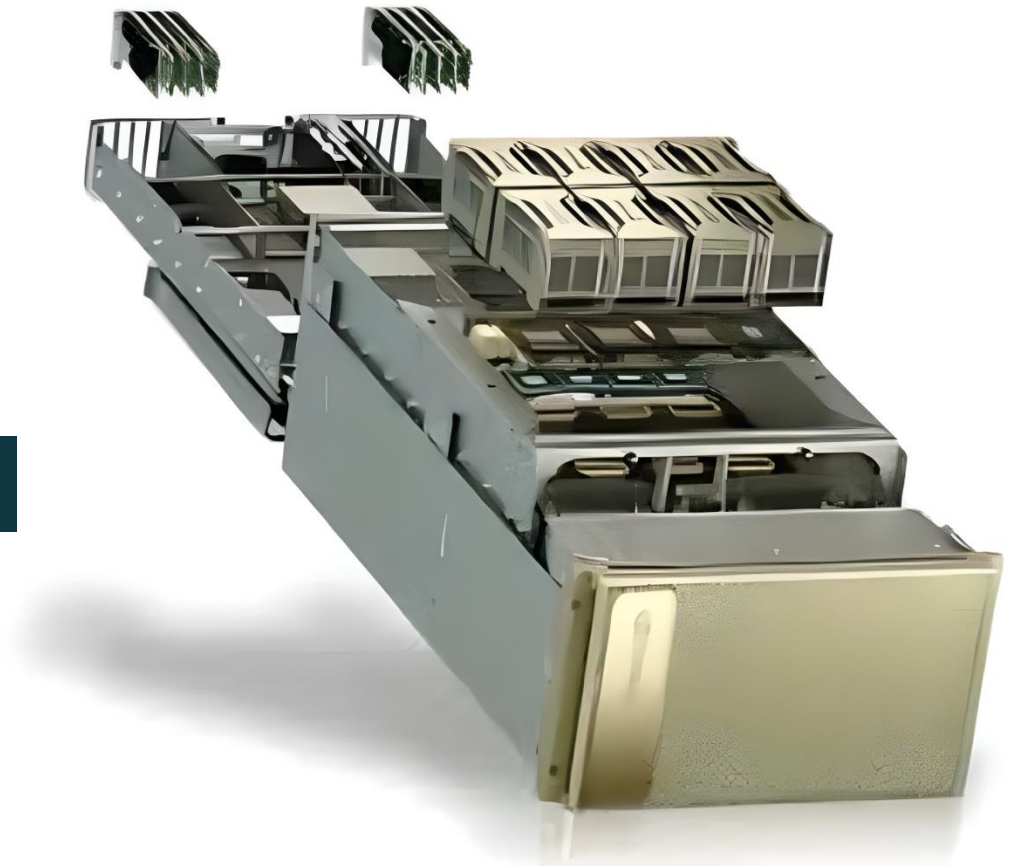
Ocean Data Processing Engine

Benchmarking

Physical hardware setup for benchmarking

Our chosen setup for the benchmark uses a general-purpose configuration (about 1 CPU / 4.7 GB RAM) suitable for most computational tasks and not optimized for any specific workload.

Fraction	Executor	Logical CPUs	Memory [GB]
Full (1×)	7	168	784
Half (½×)	4	96	448
Quarter (¼×)	2	48	224
Eighth (⅛×)	1	24	112



Nvidia DGX A100

Ocean Data Processing Engine



Datasets

Point Clouds

Status

Settings

+ New dataset

Upload Files...

Drop files here

Name	Path	Size	Format	Version	Scan	Point cloud
 HYDRO_MBES	file:/data/HYDRO2025/MBES/01-RawData/ALL	326.24 GiB	ALL	raw	Done	Done
 test_gsf005	file:/data/dataset_testing/open_source_data/GSF/MPI-M_MSM79-2/rawdata/MBES	153.4 MiB	GSF	raw	Done	Done
 test_MBES108	file:/data/dataset_testing/open_source_data/ALL/MARUM_M99/rawdata/MBES	184.0 MiB	ALL	raw	Done	No
 2025-09-09	/zeus/processing/datasets/2025-09-09/data	4.09 GiB	TIL	raw	Done	Done

Scan

Create point cloud

Reset dataset

Filename	Correctly scanned	Has valid data	EPSG code	Layers
0010_20150527_040539.all	Yes	Yes	32631	POINTS ID 255 - POINTS, POINTS ID 167 - POINTS, MBES ID 255 - MBES, MBES ID 167 - MBES
0236_20150603_122255.all	Yes	Yes	32631	POINTS ID 255 - POINTS, POINTS ID 167 - POINTS, MBES ID 255 - MBES, MBES ID 167 - MBES
0115_20150530_043930.all	Yes	Yes	32631	POINTS ID 255 - POINTS, POINTS ID 167 - POINTS, MBES ID 255 - MBES, MBES ID 167 - MBES

0115_20150530_043930.all

Number of partitions: 7

Number of elements: POINTS ID 255 - POINTS - 2946800, POINTS ID 167 - POINTS - 2946400, MBES ID 255 - MBES - 2946800, MBES ID 167 - MBES - 2946000

Benchmarking - MBES

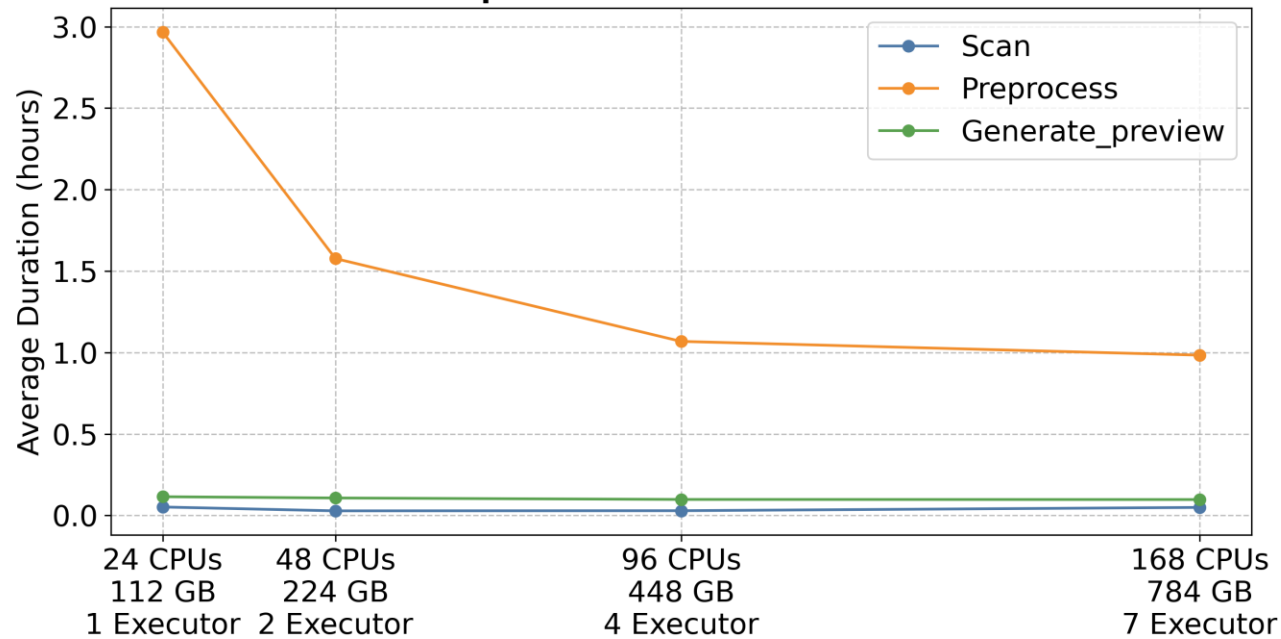
Data

- Kongsberg EM 2040 Dual Head
- 326 GB (raw data, *.all), 93 GB after preprocess
- Total number of data points **6,287,475,200**
- From initially 3 hours to 1 hour

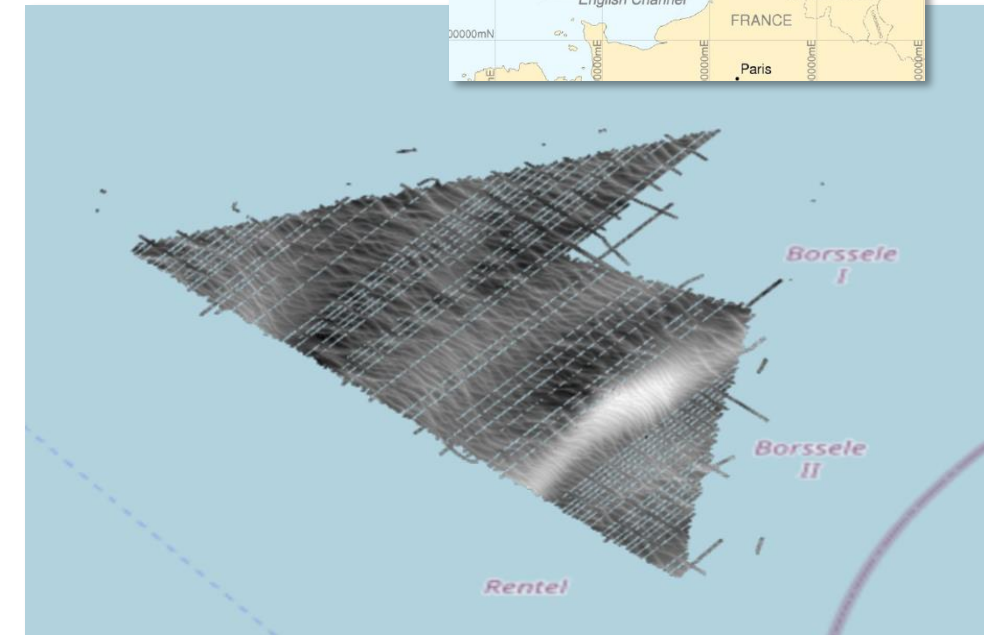
Processing Steps

- Scan
- Preprocessing (Auto-Mode)
- Preview: 1.0 m grid

Computation Times vs Resources



Resources (with total CPUs / RAM Memory)



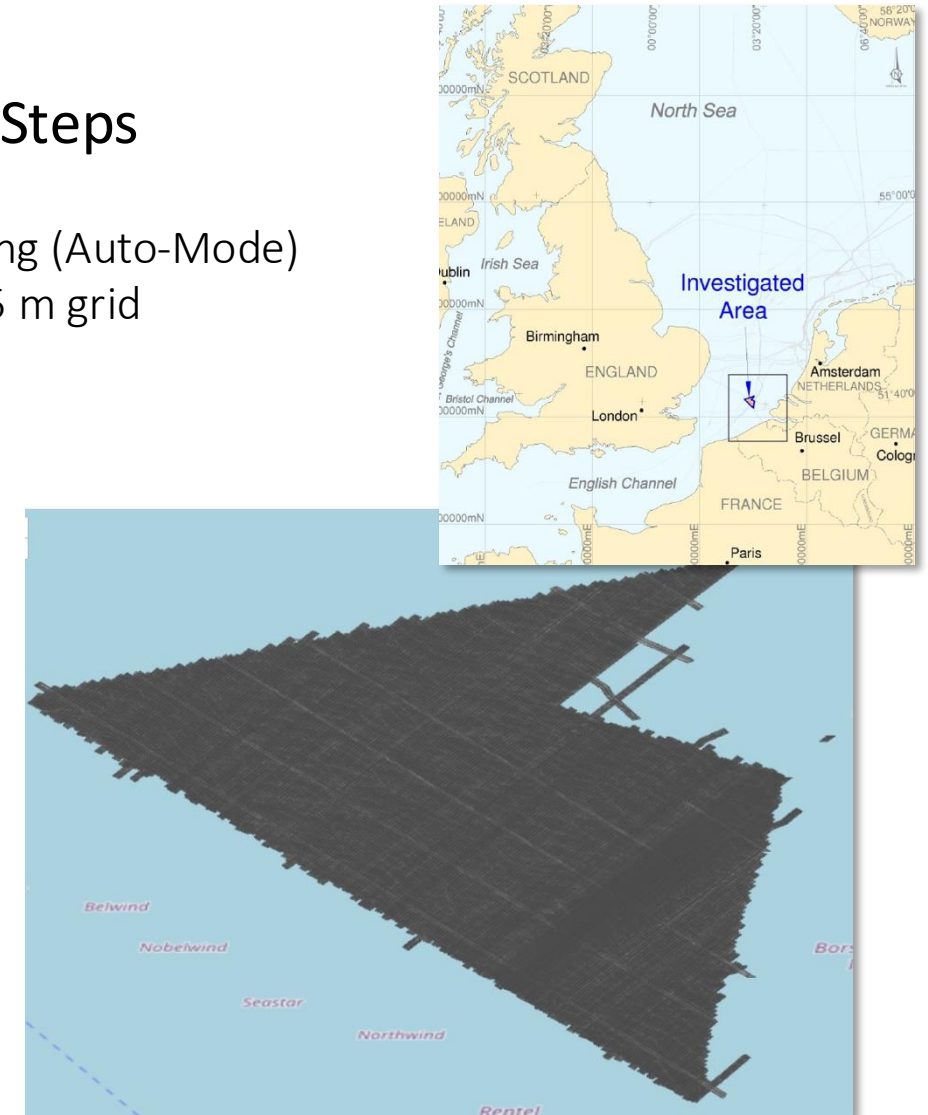
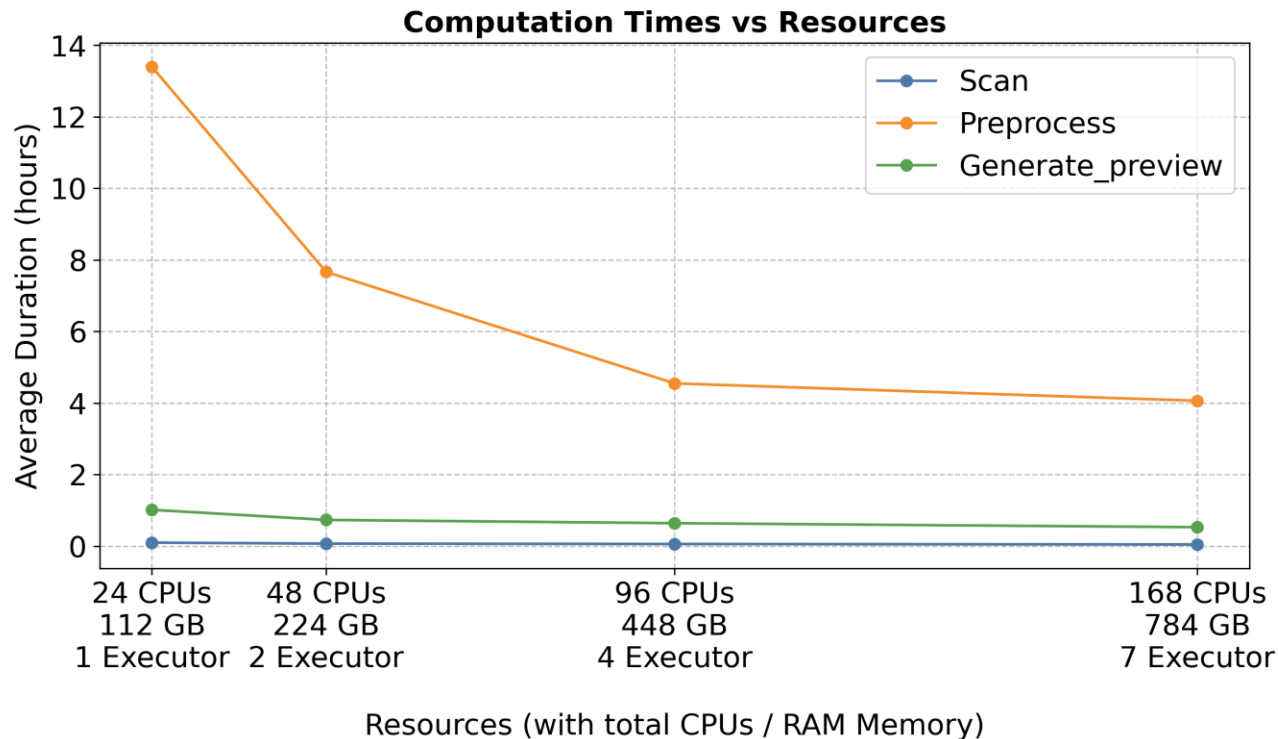
Benchmarking - SSS

Data

- EdgeTech 4200-FS
- 303 GB (raw data, *.xtf), 3.3 TB after preprocess
- Total number of data points **160,048,458,164**
- From initially 14 hours to 4 hours

Processing Steps

- Scan
- Preprocessing (Auto-Mode)
- Preview: 0.5 m grid



Parallelization

Parallelization is not only happening on the data level but also on the process level:

- **Capacity Scheduling** – based on pre-allocated resources
- **Fair Scheduling** – based on dynamic balancing and automated resource optimization



- Datasets
- Point Clouds
- Status
- Settings

Status

Command	Dataset	Started	Finished	Duration	Status	
Preprocess	SSS_Trial	2025-10-28 11:22:36	-	-	RUNNING	
Preprocess	MBES_Trial	2025-10-28 11:22:32	-	-	RUNNING	
Calculate Identifier	MBES_Demo	2025-10-28 11:22:23	-	-	RUNNING	
Calculate Gradient	SSS_Demo	2025-10-28 11:22:12	-	-	RUNNING	
Calculate Identifier	SSS_Demo	2025-10-28 11:21:26	2025-10-28 11:21:57	00:00:31	COMPLETED	
Preprocess	MBES_Demo	2025-10-28 11:20:46	2025-10-28 11:21:25	00:00:39	COMPLETED	
Scan	MBES_Demo	2025-10-28 11:20:37	2025-10-28 11:20:38	00:00:01	COMPLETED	
Preprocess	MBES_Demo	2025-10-28 11:18:58	2025-10-28 11:20:02	00:01:04	COMPLETED	

04

Use Cases

Test and technology verification with different industry groups

NATO Exercise - REPMUS

Data Fusion Cell

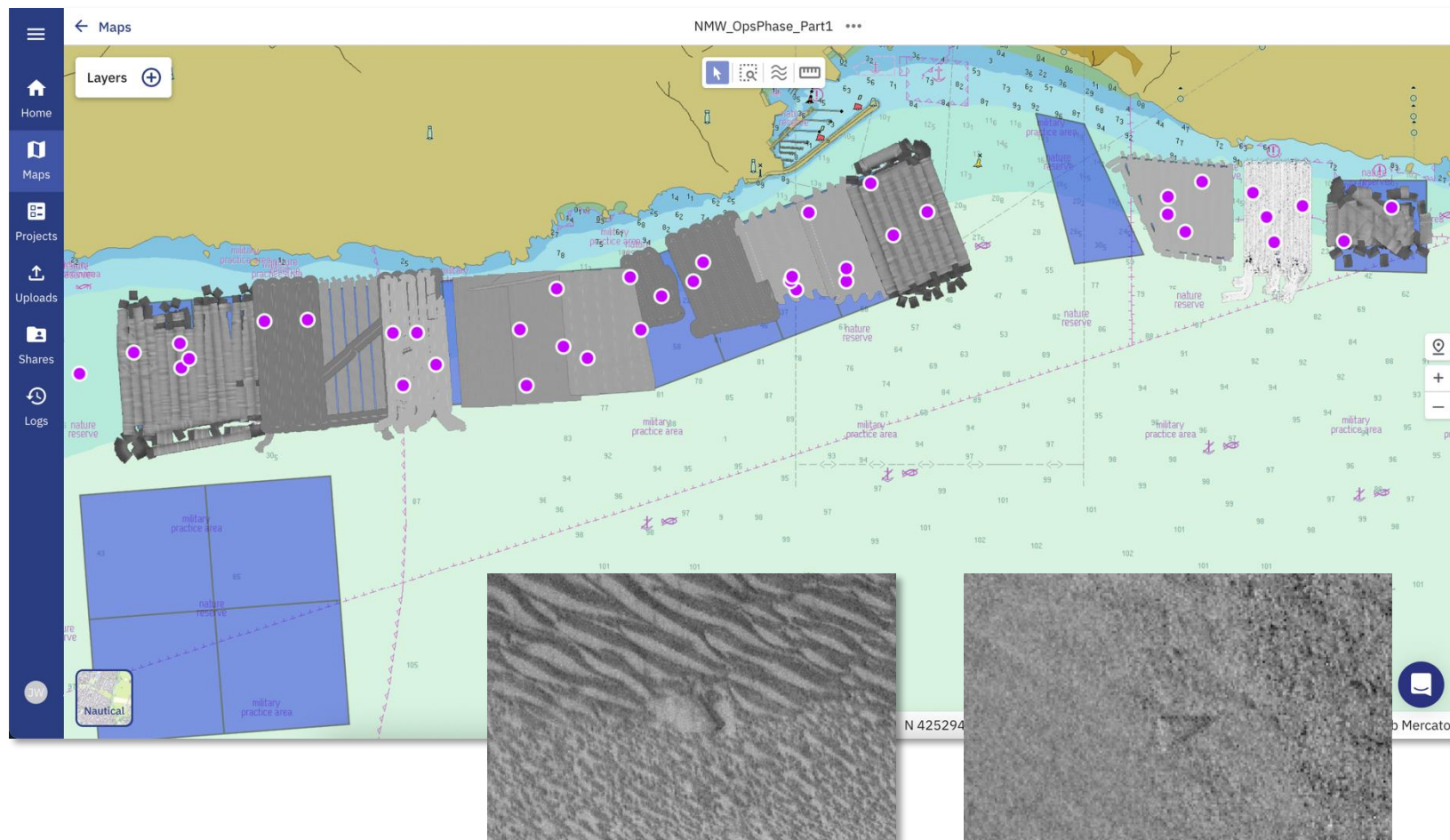
- Managing and visualizing large SSS & SAS datasets
- Processing these datasets at scale as fast as possible
- Streaming the data directly via WMS into the Common Operational Picture (COP)
- Running a fully isolated system showing the capability for classified environments
- Total: 3 TB raw data



NATO Exercise - REPMUS

Data Fusion Cell

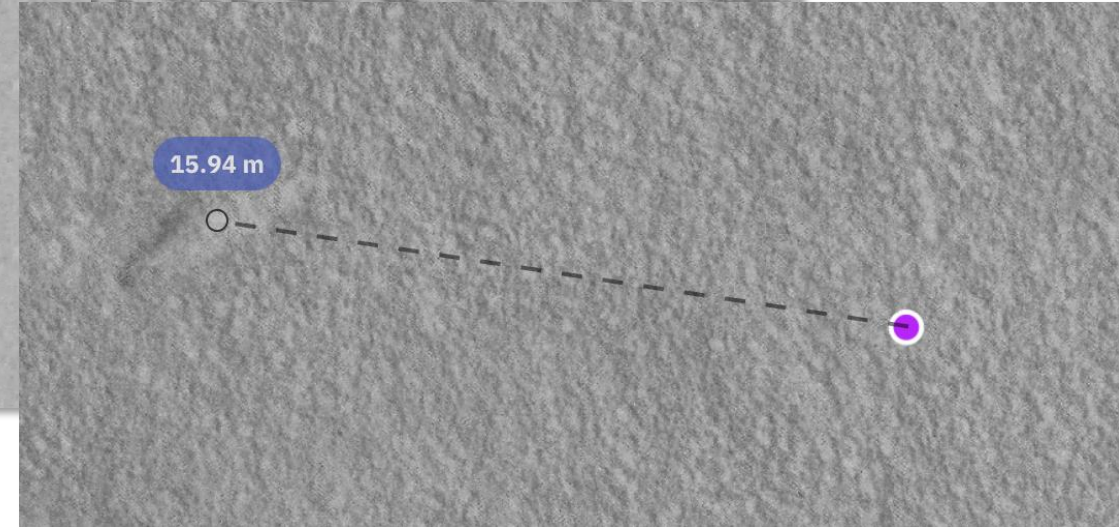
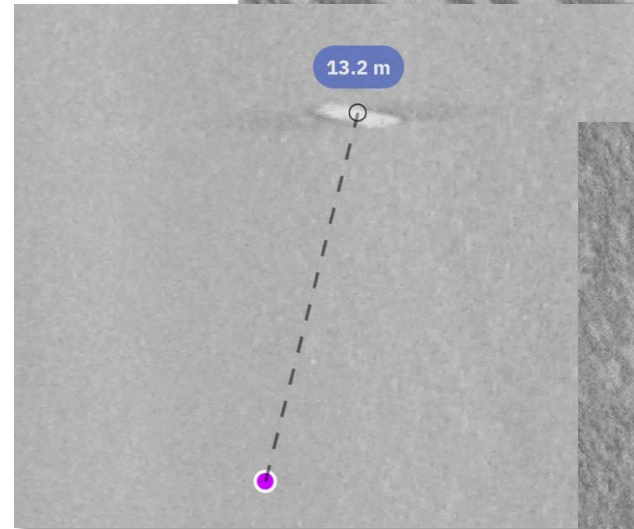
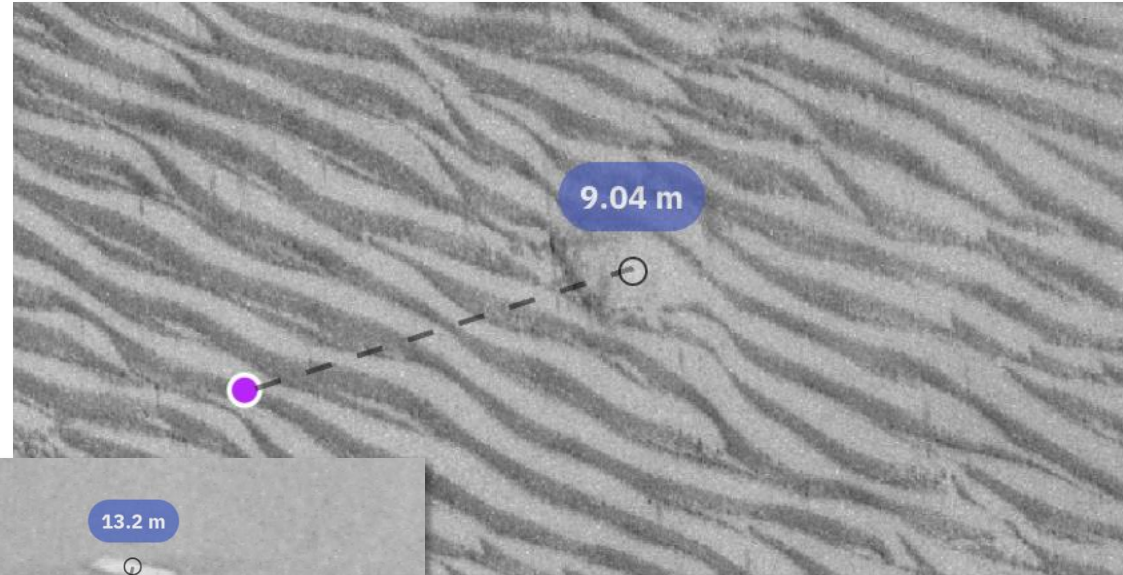
- Managing and visualizing large SSS & SAS datasets
- Processing these datasets at scale as fast as possible
- Streaming the data directly via WMS into the Common Operational Picture (COP)
- Running a fully isolated system showing the capability for classified environments
- Total: 3 TB raw data



NATO Exercise - REPMUS

Data Fusion Cell

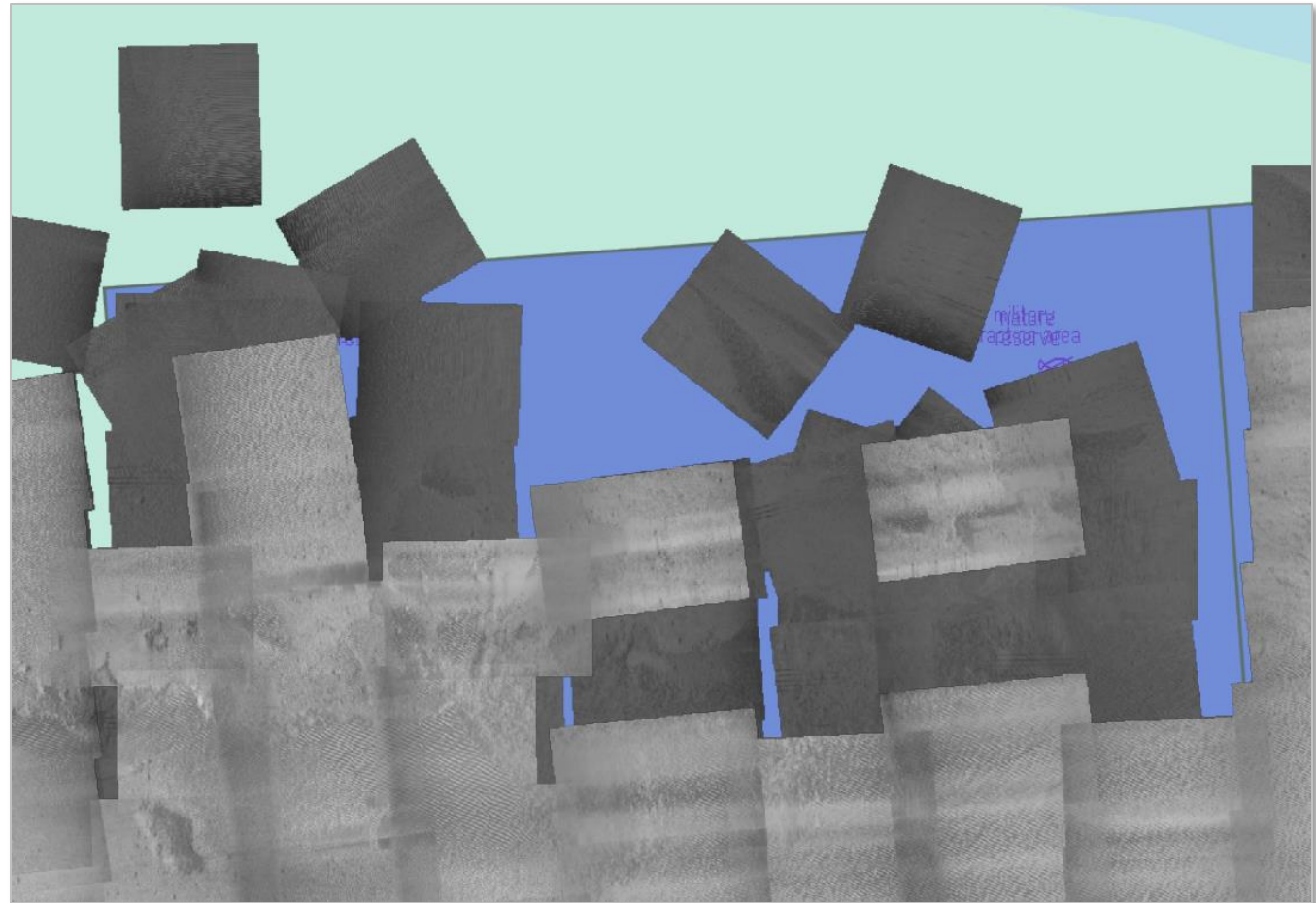
- Managing and visualizing large SSS & SAS datasets
- Processing these datasets at scale as fast as possible
- Streaming the data directly via WMS into the Common Operational Picture (COP)
- Running a fully isolated system showing the capability for classified environments
- Total: 3 TB raw data



NATO Exercise - REPMUS

Data Fusion Cell

- Managing and visualizing large SSS & SAS datasets
- Processing these datasets at scale as fast as possible
- Streaming the data directly via WMS into the Common Operational Picture (COP)
- Running a fully isolated system showing the capability for classified environments
- Total: 3 TB raw data



Benchmarking & Testing

- We are right now testing and verifying with different partners our data processing technology
- Integrating immediately feedback based on agile driven development
- Covering a broad variety of use cases, different domains and perspectives

**We are continuously looking for
interested parties testing and
verifying with us**



05

What's next?

The next steps in the technology development roadmap

- Further enhancing and automatizing the multibeam processing
- Conducting further testing on larger cloud environments: preparing a test with more than 1,000 CPUs
- Full integration of the Ocean Data Processing Engine into the Ocean Data Platform
- Continuing the development of AI-based analysis (anomaly and object detection) on pointcloud level

Contact



Jann Wendt

CEO

Mail: jwendt@north.io

Mobile: +49 1520 90 322 51

Thanks!