13TH IBERIAN GRID CONFERENCE
IBERGRID
2024
better software for better science
28-30 OCT
UNIVERSITY
OF PORTO

Contribution ID: **26** Type: **not specified**

Galician Marine Sciences Program Data Lake

Monday, 28 October 2024 17:10 (20 minutes)

Over the last two years, the Galician Marine Sciences Program (CCMM) has developed a Data Lake to support the collection and analysis of data related to Galicia's marine ecosystem. The Data Lake architecture facilitates processing both structured and unstructured data, already integrating diverse datasets such as ocean currents velocity maps, species distribution data, upwelling indices, buoy-derived marine conditions, marine carbon-related datasets, SOCAT coastal and North Atlantic data and atmospheric models.

For the storage layer, the Data Lake utilizes Apache Hadoop's HDFS distributed filesystem and Apache Parquet for efficient distributed and parallel processing.

For the analysis layer, Apache Spark enables high-performance, scalable data processing, combining multiple datasets to advance marine ecosystem research and support sustainable resource management.

Interactive processing is enabled through a web portal that uses JupyterLab notebooks tightly integrated with the Data Lake and customized for marine sciences usage.

The Data Lake not only accelerates data-driven insights but also provides a scalable infrastructure for future research, fostering collaboration and innovation in the sustainable management of Galicia's marine resources.

- https://ccmmbigdata.cesga.es/
- https://ccmmbigdata.cesga.es/datasets
- https://cienciasmariñas.gal/lang=en

Primary authors: Ms GRELA LLERENA, Cecilia (CESGA); Mr PRIETO RÚA, Pablo (CESGA); Dr CACHEIRO LÓPEZ, Javier (CESGA); Dr FERNANDEZ SANCHEZ, Carlos (CESGA); Mr LANDEIRA VEGA, Francisco (CESGA)

Presenters: Ms GRELA LLERENA, Cecilia (CESGA); Mr PRIETO RÚA, Pablo (CESGA); Dr CACHEIRO LÓPEZ, Javier (CESGA)

Session Classification: IBERGRID

Track Classification: Development of innovative software and services