KER 5 | URV

Dataplug

TYPE

[SERV] [METH]

TECHNICAL READINESS LEVEL

TRL6

INTELLECTUAL PROPERTY RIGHTS

COPYRIGHT

EXPLOITATION ROUTE

OPEN SOURCE

Dataplug is a client-side only, extensible, Python framework with the goal of enabling efficient data partitioning of unstructured scientific data stored in object storage (like Amazon S3) for elastic workloads. Dataplug enables on-the-fly partitioning for supported data formats, driving important savings in data ingestion stage.

KEY BENEFITS FOR COMPUTE CONTINUUM PROJECTS

The entire compute continuum counts with heterogeneous resources with highly variable characteristics (e.g., computational power). Dataplug enables data partitioning tailored to this complexity: each worker (running on cloud, edge, or HPC) will only consume the necessary data chunk directly from the source, increasing abstraction in data ingestion across the compute continuum.

FOR RESEARCHERS AND INDUSTRIAL TEAMS:

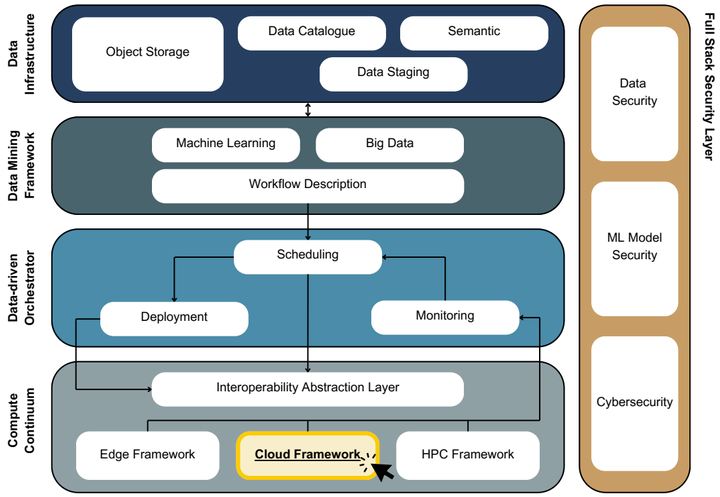
- Dataplug dynamic partitioning frees users from fixed static partitions. Only one preprocessing stage is required.
- Dataplug is perfect for optimizing resource provisioning: allowing for the fine-tuning of computational resources by launching with different partition sizes.

FOR ADMINISTRATORS AND CDOS

- Dataplug avoids fixed and sometimes slow stages of static partitioning.
- Dataplug is very-easy to use and to integrate in different frameworks (e.g., Lithops).

USE AND IMPACT BEYOND EXTRACT PROJECT AND ITS PARTNERS

Dataplug has expanded the different data formats supported since the beginning of the EXTRACT project. Dataplug now includes an implementation of Measurement Sets (main source of TASKA-C pipeline). Dataplug formats may be incremented in the future, with greater supportability for data formats each time.



https://gitlab.bsc.es/extract/extract-sa/lithops

This project has received funding from the European Union's Horizon Europe programme under grant agreement number 101093110.