Data ETL Tasks for a Data Engineer

Owner	Leo Li
Tags	
Last edited time	@March 1, 2024 2:46 PM

Task 1: Converting NetCDF Files to ZARR Format

We are interested in ingesting the IMOS Satellite Remote Sensing dataset from 2023 to the most recent date available. The source files are available at: https://thredds.aodn.org.au/thredds/catalog/IMOS/SRS/OC/gridded/aqua/P1D/catalog.html.

Your task is to ingest the dataset files with the `.aust.ipar.nc` suffix. An example includes the first two days' source files of 2023, available at https://thredds.aodn.org.au/thredds/catalog/IMOS/SRS/OC/gridded/aqua/P1D/2023/01/catalog.html.

Catalog https://thredds.aodn.org.au/thredds/catalog/IMOS/SRS/OC/gridded/aqua/P1D/2023/01/catalog.html

Dataset	Size	Last Modified
01		--
A.P1D.20230101T053000Z.aust.K_490.nc	20.13 Mbytes	2023-04-27T23:20:16Z
A.P1D.20230101T053000Z.aust.chl_gsm.nc	23.07 Mbytes	2023-04-27T23:19:57Z
A.P1D.20230101T053000Z.aust.chl_oc3.nc	23.14 Mbytes	2023-04-27T23:20:00Z
A.P1D.20230101T053000Z.aust.chl_oci.nc	29.43 Mbytes	2023-04-27T23:20:04Z
A.P1D.20230101T053000Z.aust.dt.nc	6.295 Mbytes	2023-04-27T23:20:08Z
A.P1D.20230101T053000Z.aust.ipar.nc	18.52 Mbytes	2023-04-27T23:20:12Z
A.P1D.20230101T053000Z.aust.l2_flags.nc	6.500 Mbytes	2023-04-27T23:20:20Z
A.P1D.20230101T053000Z.aust.nanop_brewin2012in.nc	6.700 Mbytes	2023-04-27T23:20:24Z
A.P1D.20230101T053000Z.aust.npp_vgpm_eppley_gsm.nc	9.517 Mbytes	2023-04-27T23:20:27Z
A.P1D.20230101T053000Z.aust.npp_vgpm_eppley_oc3.nc	9.413 Mbytes	2023-04-27T23:20:31Z
A.P1D.20230101T053000Z.aust.par.nc	11.86 Mbytes	2023-04-27T23:20:39Z
A.P1D.20230101T053000Z.aust.picop_brewin2012in.nc	6.637 Mbytes	2023-04-27T23:20:42Z
A.P1D.20230101T053000Z.aust.sst.nc	15.80 Mbytes	2023-04-27T23:20:46Z
A.P1D.20230101T053000Z.aust.sst_quality.nc	3.350 Mbytes	2023-04-27T23:20:50Z
A.P1D.20230102T053000Z.aust.K_490.nc	23.41 Mbytes	2023-04-27T23:21:18Z
A.P1D.20230102T053000Z.aust.chl_gsm.nc	27.07 Mbytes	2023-04-27T23:20:58Z
A.P1D.20230102T053000Z.aust.chl_oc3.nc	27.10 Mbytes	2023-04-27T23:21:02Z
A.P1D.20230102T053000Z.aust.chl_oci.nc	33.97 Mbytes	2023-04-27T23:21:06Z
A.P1D.20230102T053000Z.aust.dt.nc	6.862 Mbytes	2023-04-27T23:21:10Z
A.P1D.20230102T053000Z.aust.ipar.nc	21.67 Mbytes	2023-04-27T23:21:14Z
A.P1D.20230102T053000Z.aust.l2_flags.nc	7.112 Mbytes	2023-04-27T23:21:22Z
A.P1D.20230102T053000Z.aust.nanop_brewin2012in.nc	8.883 Mbytes	2023-04-27T23:21:25Z
A.P1D.20230102T053000Z.aust.npp_vgpm_eppley_gsm.nc	13.51 Mbytes	2023-04-27T23:21:29Z
A.P1D.20230102T053000Z.aust.npp_vgpm_eppley_oc3.nc	13.38 Mbytes	2023-04-27T23:21:33Z
A.P1D.20230102T053000Z.aust.par.nc	15.85 Mbytes	2023-04-27T23:21:41Z
A.P1D.20230102T053000Z.aust.picop_brewin2012in.nc	8.799 Mbytes	2023-04-27T23:21:45Z

The goal is to ingest the dataset from the first day of 2023 and convert the NetCDF files into the ZARR format (details on ZARR can be found here: https://wiki.earthdata.nasa.gov/display/ESO/Zarr+Format). You can use any tools or methods for the conversion. What would be your chunking strategy to make the resulting file more efficiently accessed?

Note: This dataset is updated daily. If you would like to automate the ZARR conversion to update daily, what will be your strategy?

Task 2: Converting CSV to GeoParquet

We are also interested in a CSV file from the Australian Bureau of Statistics named `abs-regional-lga-2021`. It contains several demographic indicators in each of the local government areas (LGA). The file can be downloaded through this URL: https://api.data.abs.gov.au/files/ABS_ABS_REGIONAL_LGA2021_1.0.0.csv.

More information about the source data can be found here: https://www.abs.gov.au/methodologies/data-region-methodology/2011-23.

Your task is to ingest the `abs-regional-lga-2021` dataset and convert it into the GeoParquet format. The desired outcome should resemble this: s3://gbr-dms-data-public/abs-regional-lga-2021/data.parquet (Accessing the result by AWS s3). You may need to clean the source CSV, pivot the table, and add a geometry column for this dataset. The reference geometry is accessible programmatically at s3://gbr-dms-data-public/tasks/geoms.parquet. We prefer you use Python for this task.

What would be your partition strategy for making the access of the resulting GeoParquet file more efficient?

If you have any questions about the tasks, please write to Leo.Li@utas.edu.au