This temporary repo stores dask examples that were developed for NCI’s training purpose. They will be released through NCI’s training site after a period of peer-review process.

Notes:

1. Those examples are in dev mode, hence might not be perfect such as comments, structure and design, etc. Feedbacks are surely welcome to help improving the quality.
2. I have been preparing more examples, will upload them as soon as I complete them. Therefore, you might see something different when cloning the repo each time.
3. A few packages will need to installed: graphviz, holoviews, datashader, pyarrow, bottleneck etc.
4. Typos might not be checked throughout
5. Some examples (especially the toy problem ones) are copied from dask tutorial git reop, which are referenced inside of the examples.
6. Create cluster like the cell below is not necessary when running some of the examples. It was designed to demonstrate distributed calculation through dashboard when running those examples in Pangeo environment on Gadi. You can decide whether to use it at your own interests.

*From dask.distributed import Client,LocalCluster*

*client = Client(scheduler\_file=’scheduler.json’)*

*print(client)*

|  |  |  |
| --- | --- | --- |
| Example | datasets | description |
| 1\_11\_basics | toy | Dask array, lazy loading and graph visualisation, progress bar, etc |
| 1\_12\_chunks | CMIP6 – oi10 | Read in CMIP6 by giving a chunksize |
| 1\_13\_delayed\_toy | toy | Dask.delayed utility, apply it to a loop example to accelerate |
| 1\_14\_delayed\_pandas\_paleoceanography | Paleo climate csv | Apply dask.delayed to paleoclimate records (in csv files)  Notes: Running this example on local computer is fine, but not on HPC with csv files not found error. I found [a similar issue](https://stackoverflow.com/questions/50987030/file-not-found-error-in-dask-program-run-on-cluster) and it may explain why and work around are discussed. |
| 1\_15\_dataframes\_weather station | rainfall csv | Readin BOM’s rainfall csv data as dask.dataframe, and loop over a number of csv files (as chunks) to demonstrate distributed graph calculation  Read/Write to Parquet for a better performance |
| 1\_16\_schedulars\_weather station | rainfall csv | Apply Dask Schedulars (local thread,  local processes, single thread, distributed schedulars)to the same workflow above. |
| 1\_17\_Numpy\_AGDC  Note: The first little session needs a large HDF5 file. Given the git limits on file size, I can not make it available, but I can make it available on gadi:/scratch/public/jbw900/data/  the last line, save to hdf5 does not work. | Gridded precipitation – rr8 | Synthetic data to demonstrate Dask Array (similar to NumPy). Explain parallelism (by chunks) on array for better performance by apply to AGDC datasets。 |
| 1\_18\_DataArray\_Xarray\_CMIP6  Note: Sometimes, I got killworker msg. | CMIP6 – fs38 | Dask.dataArray vs. Xarray.dataArray with the same standard xarray operations. |
| 1\_19\_Xarray\_Gridded\_Percipitation | CMIP6 – fs38 | Analysis of Gridded Ensemble Precipitation Estimates using CMIP6 model output. Extract a time series of annual maximum precipitation events over a region |
| 1\_20\_ ocean model\_analysis\_visulisation | CMIP6 – oi10 – GFDL highresSST-future | Using Holoviews and Datashader to visualise data  Known issue: Datashader shows snapshot very slowly |