How to handle large model output

ACCESS Training Day 2024

Australia's climate simulator

ACCESS
National Research Infrastructure

What's so hard about processing large datasets?

What's so hard about processing large datasets?

Common problems

Data is bigger than available memory

😫

Takes a really long time to process

What's so hard about processing large datasets?

Common problems

Data is bigger than available memory

Takes a really long time to process

Helpful strategies

Parallel & distributed computing

Read and process data in chunks

Lazy loading/evaluation

Out-of-core computation

What's so hard about processing large datasets?

Common problems

Data is bigger than available memory

Takes a really long time to process

Helpful strategies

Parallel & distributed computing

Read and process data in chunks

Lazy loading/evaluation

Out-of-core computation

Useful software

xarray

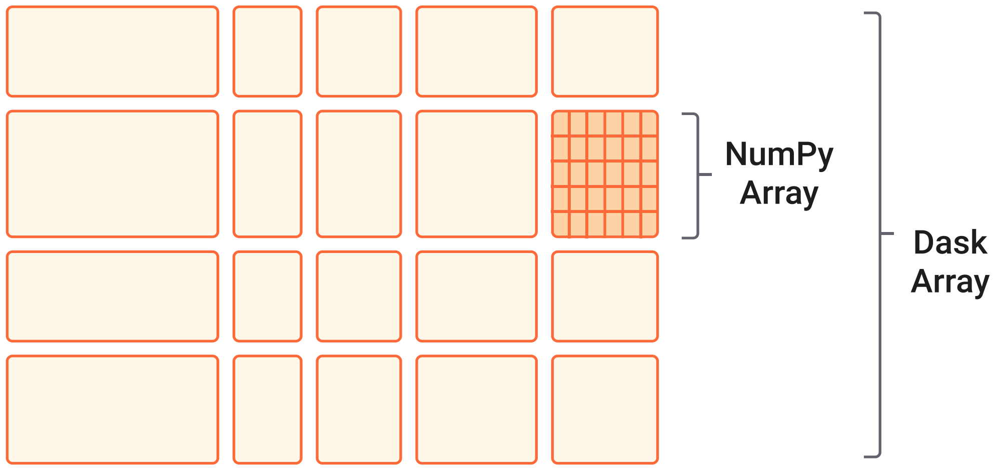
dask

What is Xarray and what is Dask?

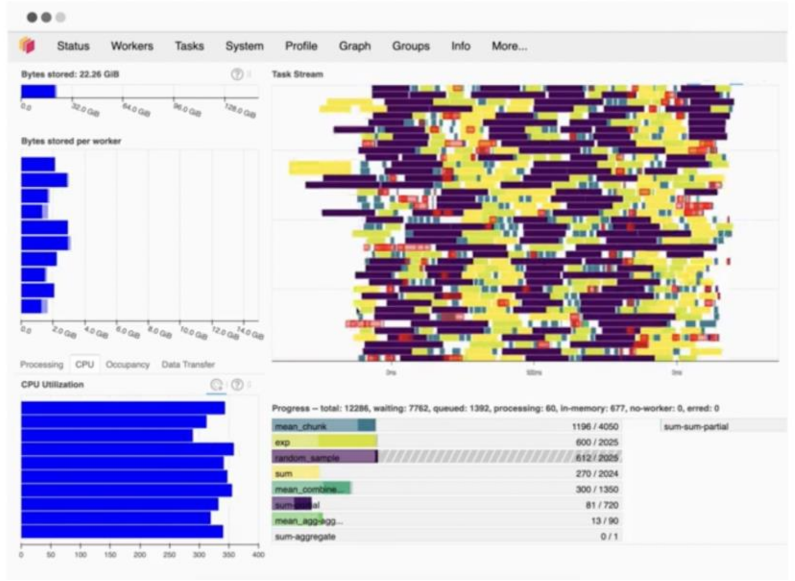
dask

Python library for parallel and distributed computing

Dask Arrays



Dask Computation
(Dask workers & Dask scheduler)

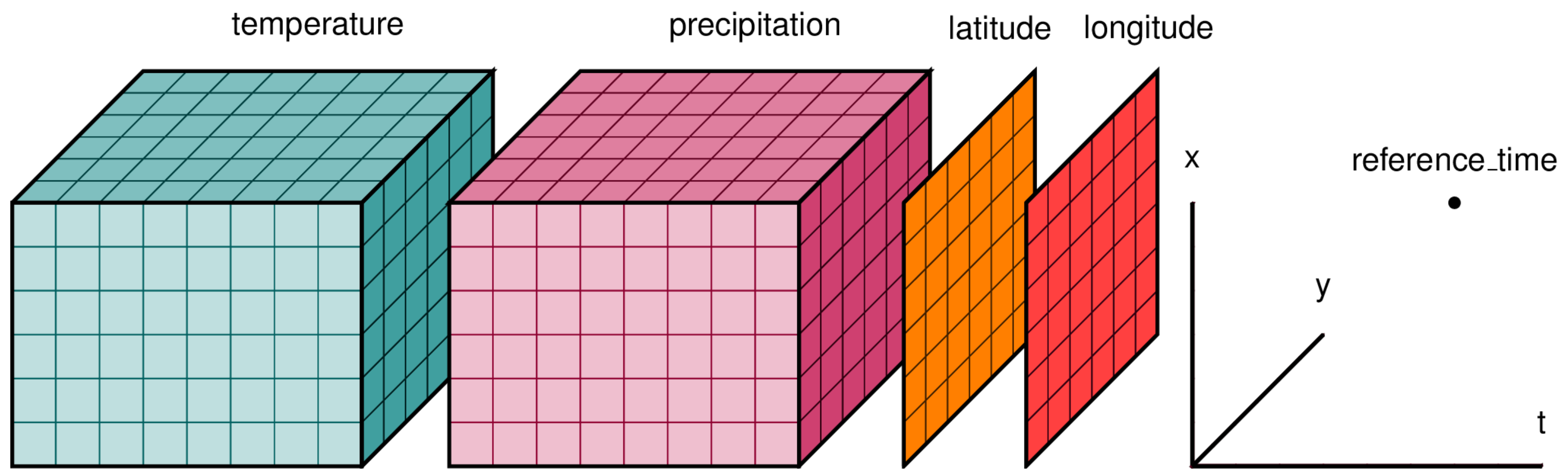


Progress -- total: 12286, waiting: 7762, queued: 1392, processing: 60, in-memory: 677, no-worker: 0, erred: 0

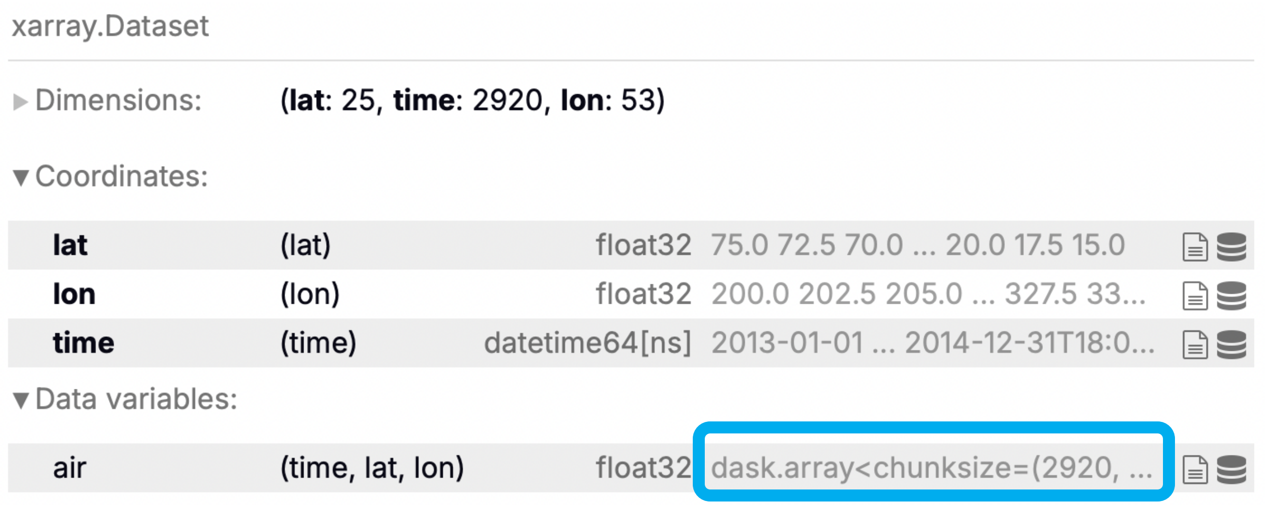
mean_chunk	1196 / 4050	sum-sum-partial
exp	600 / 2025	
random_sample	612 / 2025	
sum	270 / 2024	
mean_combine...	300 / 1350	
sum-partial	81 / 720	
mean_agg-agg...	13 / 90	
sum-aggregate	0 / 1	

What is Xarray and what is Dask?

xarray N-D labeled arrays and datasets in Python



Supports chunking and lazy loading by wrapping Dask Arrays



xarray.Dataset

▸ Dimensions: (**lat**: 25, **time**: 2920, **lon**: 53)

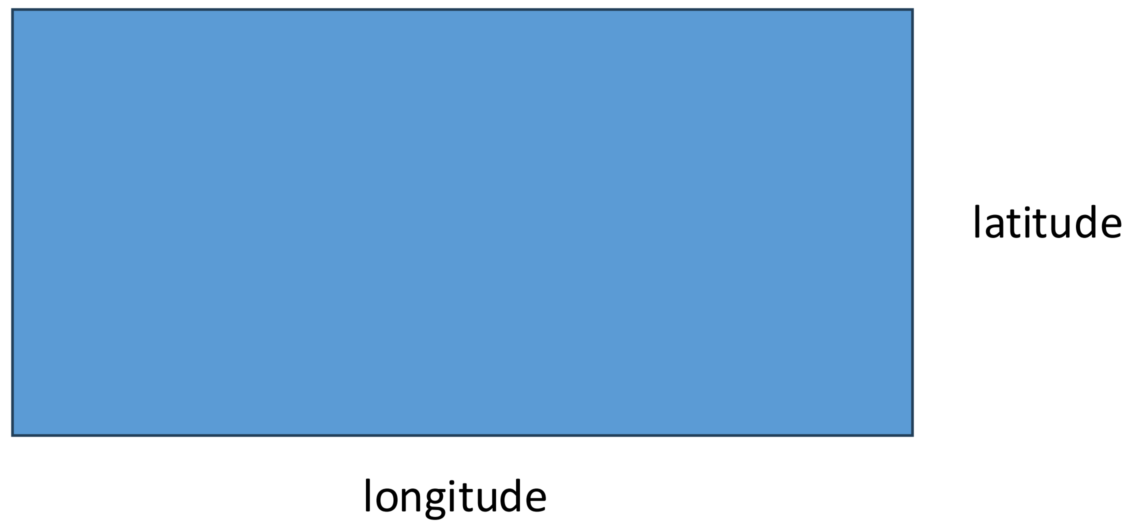
▾ Coordinates:

lat	(lat)	float32	75.0 72.5 70.0 ... 20.0 17.5 15.0
lon	(lon)	float32	200.0 202.5 205.0 ... 327.5 33...
time	(time)	datetime64[ns]	2013-01-01 ... 2014-12-31T18:0...

▾ Data variables:

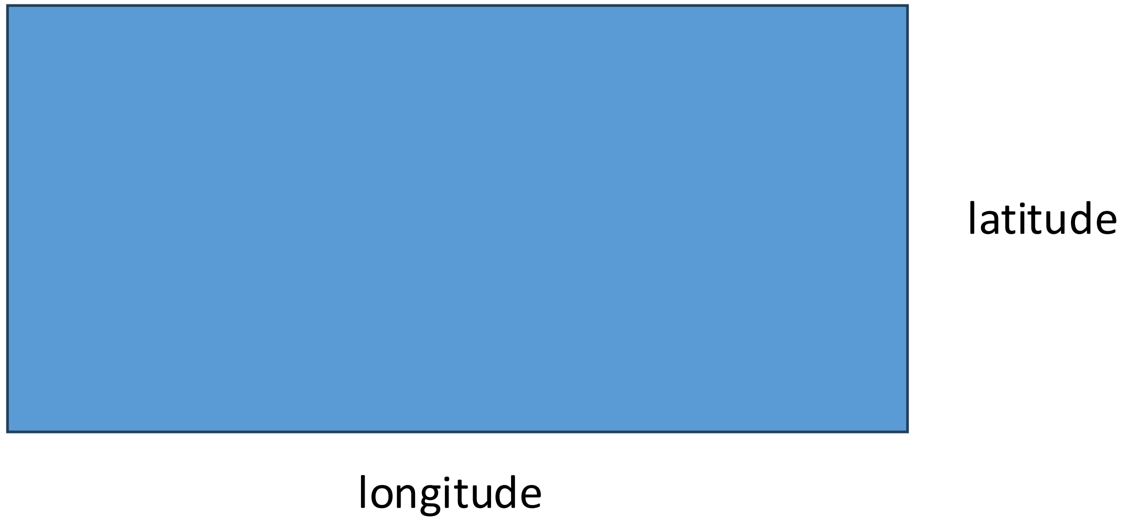
air	(time, lat, lon)	float32	dask.array<chunksize=(2920, ...

Chunked data

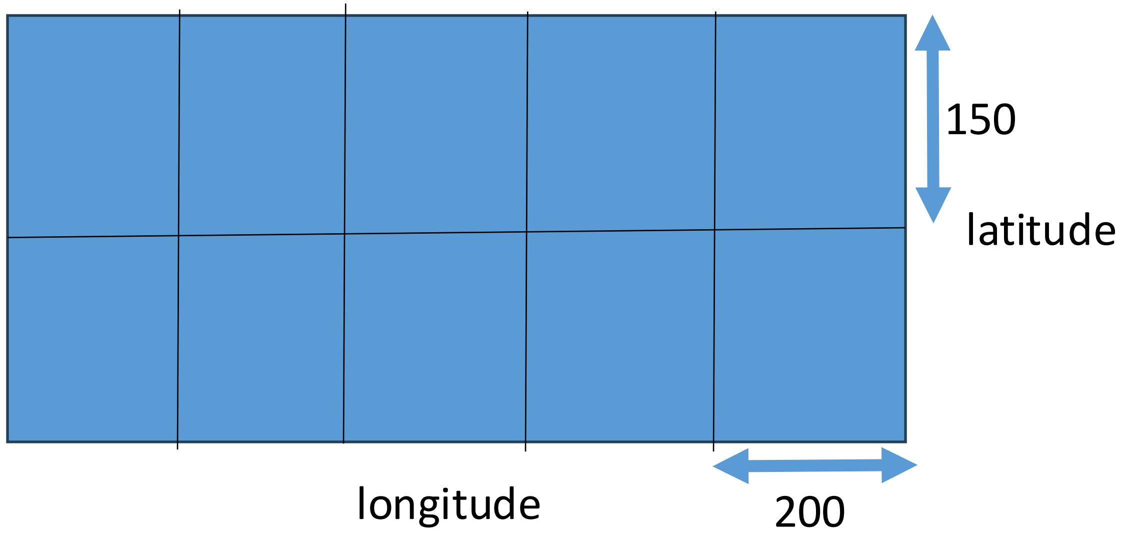


(1000,300)

Chunked data



(1000,300)

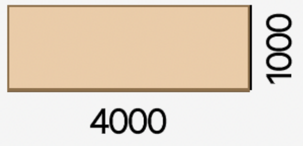


10 chunks
(200,150)

Dask and chunked data

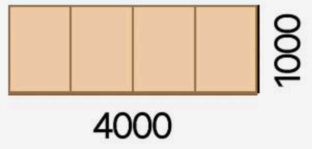
```python
import dask.array as da
ones = da.ones(shape)
ones
```

	Array	Chunk
Bytes	30.52 MiB	30.52 MiB
Shape	(1000, 4000)	(1000, 4000)
Count	1 Tasks	1 Chunks
Type	float64	numpy.ndarray



```python
chunk_shape = (1000, 1000)
ones = da.ones(shape, chunks=chunk_shape)
ones
```

	Array	Chunk
Bytes	30.52 MiB	7.63 MiB
Shape	(1000, 4000)	(1000, 1000)
Count	4 Tasks	4 Chunks
Type	float64	numpy.ndarray



Activity 1:
How does chunking + parallel computing speed up computations?

ACCESS
National Research Infrastructure

Activity 1:
How does chunking + parallel computing speed up computations?

Lesson learned: running computations on chunks in parallel speeds up the computation

Activity 2a:
How to best choose a chunk size?
Too ***small*** can be problematic!

ACCESS
National Research Infrastructure

Activity 2a:
How to best choose a chunk size?
Too ***small*** can be problematic!

Lesson learned: too small chunk sizes creates large overhead for scheduler and results in slow computation

Activity 2b:
How to best choose a chunk size?
Too ***big*** can also be problematic!

ACCESS
National Research Infrastructure

Activity 2b:
How to best choose a chunk size?
Too ***big*** can also be problematic!

Lesson learned: too big chunk sizes can cause memory errors and prevent the computation from finishing

Activity 3:
How to decide which dimensions to chunk along?

Activity 3:
How to decide which dimensions to chunk along?

Lesson learned: chunked dimensions must make sense for your computation

ACCESS
National Research Infrastructure

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

Australia's climate simulator