Lecture 4:
[Optional] Dask for parallel and distributed processing

Kosta Rozen

When data is huge

Apache Spark: distributed analytics engine

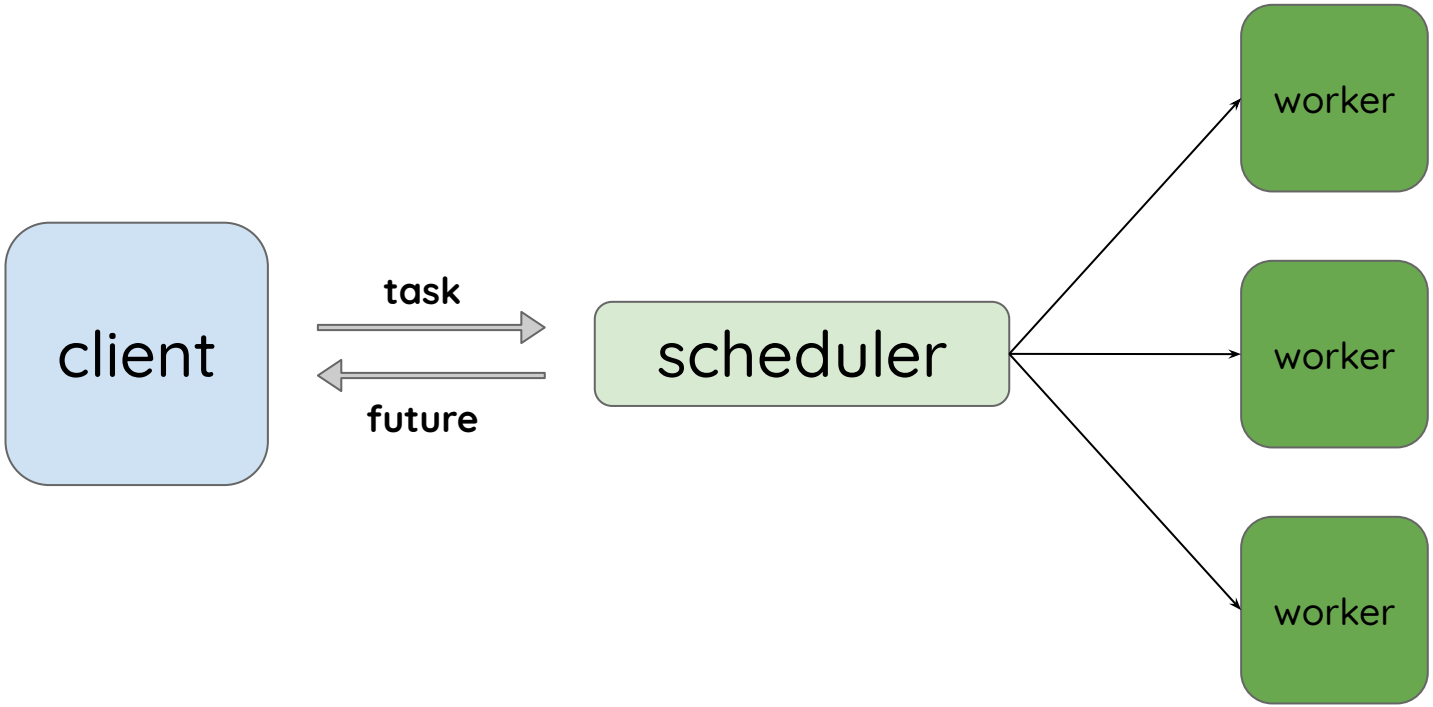
- in memory
- can handle streaming jobs
- knows about ML
- and graph data

Dask

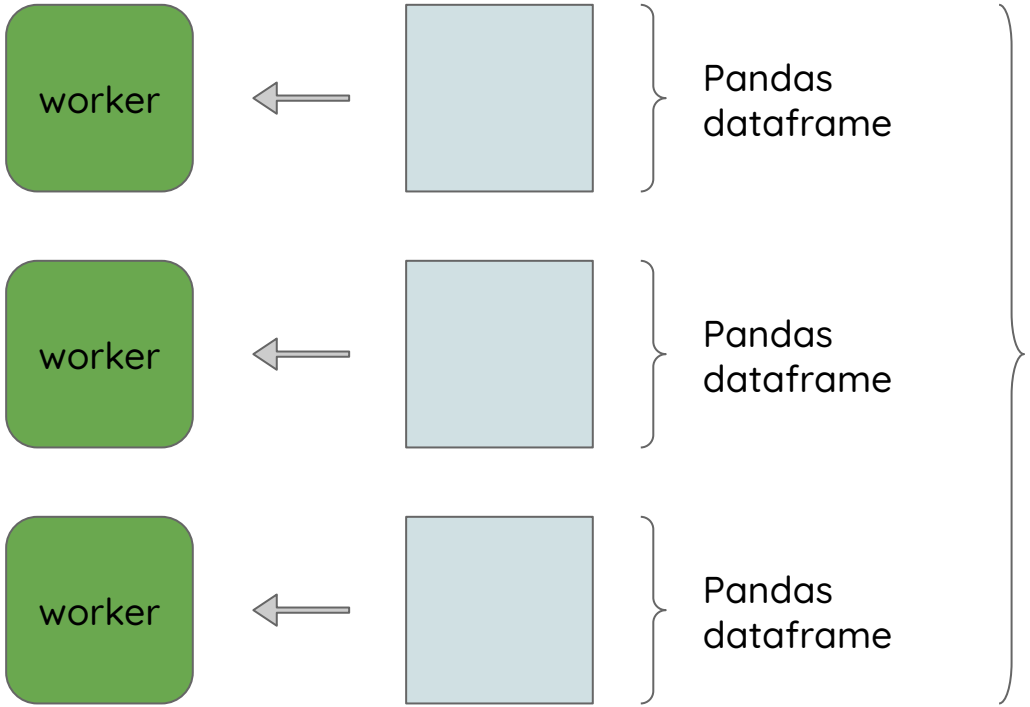
But can we go with Python?

- Dask is (to some extent) a replacement to Spark (at least up to tens of Tb)
- very easy to setup and experiment (even locally)
- flexible containers (distributed dataframes, arrays, bags)
- some operations (notably - joins on non-index columns) are still costly and may benefit from manual tweaking

Dask cluster



Dask dataframes



Dask dataframe
(split by index, index-based join's and groupby's are ok)