You can use Dask dataframes instead of pandas to read the parquet files without loading them into memory.

Example: Read parquet, only keep every 4th era, and compute to pandas

import dask.dataframe as dd training_data = dd.read_parquet('train.parquet')

pare down the number of eras to every 4th era

every_4th_era = training_data[ERA_COL].unique()[::4].compute() training_data = training_data[training_data[ERA_COL].isin(every_4th_era)].compute()