ParallelProcessing

Christian Torp-Pedersen

2023-07-20

This document is specifically focused on speeding up data management of files in Statistics Denmark for the many cases that data are stored in multiple files, typically one file per year. The code uses the BEF-data as example. These files include several basic pieces of information for each person and for each year. The programs below will extract limited information from each file and output a single file with the extracted data.

Basic information for people in Statistics Denmark is kept in BEF once yearly and from 2008 once for each quarter.

Whether parallellisation is achived with SAS or R it is important to realize that each of the jobs defined to be run in parallel function independently. Therefore additional information has to be passed to the process. SAS libnames will be unknown to the new process unless it has specifically been provided. Similarly all packages needed for each proces has to be provided in R.

Using R

```
library(heaven)
library(data.table)
library(doParallel)
library(foreach)
# Create a filelist with those of interest. In this case a regular
# expression is used to select file starting with 'bef200' and followed
# by 4-6
filelist <- list.files('X:/..relevant direc..','^bef2000[6-8].*',full.names = TRUE)
#Made a cluster for calculationt, this one with 10 cores:
cl <- makeCluster(10)
birthSex <- rbindlist(#combines the results to a single dataset
  foreach(x=1:length(filelist),.packages("heaven","data.table")) %dopar% {
    dat <- importSAS(filelist[x],keep("pnr","koen","foed_dag"))
  })
stopCluster(cl)
gc() # Always good to run the garbage collecter and free memory!
```

Using SAS

```
libname mydata 'Directory with bef...';
libname temp 'My temporary directory';
options threads cpucount=10;
options autosignon;
%let last_year_bef=2023; *only relevant if this year is included;
%let last_quarter=2; *The last year in DST only has som quarters;
```

```
%let lst= 03 06 09 12;
%let endloop=4;
%macro agesex;
 %do i=2006 %to 2008;
    %if &i=&last_year_bef %then %let enddloop=&last_quarter; *not relevant for example;
    %do ii=1 %to &endloop;
      %syslput _local_/remote=t&i&ii;
      rsubmit t&i wait=no connectpersist=no inheritlib=(temp mydata)
        data temp.bef&i.%scan(&lst,&ii); set mydata.bef&i.%scan(&lst,&ii)(keep pnr koen foed_dag);
      endrsubmit;

 waitfor _all_;
 signoff _all_;
%mend;
%agesex;
data mineSamleData; set temp.bef:;  run; *alle filer samles i "work";
proc datasets nolist lib=temp;
 delete bef:; *clear up!;
run;
```