Subject: Can DW handle big data?
Posted by greatzdl on Tue, 22 Oct 2019 06:00:54 GMT
View Forum Message <> Reply to Message

Dear Thomas,
I have used datawarrior for many years since it was released as open source program. Thank you very much for your contibution.

During these years, I found that it is hard to open data with rows larger than 3M, especially with structure columns. Do you have any solutions to solve this problem? DO you have plans to use multithreading technology to open large data file?

Hope to get your relpy.

Best wishes

DaRong

Subject: Re: Can DW handle big data?
Posted by thomas on Thu, 24 Oct 2019 18:59:06 GMT
View Forum Message <> Reply to Message

Dear DaRong,

3 million rows is already a lot. I recommend for very large files to use DataWarrior on Linux, because you can easily increase the memory maximum that there is at least no memory problem. DataWarrior uses multithreading for most functions, which benefit from it. However, reading a file is a serial process and cannot easily be parallelized. Possibly I could gain some performance, when distributing the data analysis after file loading on multiple cores. I will put it on the agenda, but not before the next release, which I anticipate before the end of the year.

Thanks and best wishes,

Thomas

Subject: Re: Can DW handle big data?
Posted by nbehrnd on Sun, 27 Oct 2019 16:01:40 GMT
View Forum Message <> Reply to Message

In addition, DaRong, if working in Linux and facing limitation by the RAM accessible on your computer, you may supplement «working memory» with a swap partition. While it won't be as performant in terms of read-write access speed, especially if it is on a HDD platter, as a true RAM brick, this offers a noticable benefit quickly setup (e.g. using an Ubuntu session on AWS).

A possible primer may be
https://linuxize.com/post/how-to-add-swap-space-on-ubuntu-18 -04/

Norwid