Loading a Huge CSV File

If the job uses a file: URL, the job engine just uses it as-is. However, if the job uses a repository: URL, then the job engine needs to copy the entire file to the local tmp directory so it has a file: URL. (the job engine does not know the machine(s) on which dacapo has its store).

The following steps happen in the background:

1. The data service receives a request from the Designer and creates a job in the job engine.

2. The job engine fetches the xxxMB from DaCapo into a local folder (slow).

3. The job engine turns the CSV into XML in a local folder (slow).

4. The job engine writes the XML back into DaCapo /Temp/<user>/tmp-0000XXX.xml (slow).

5. The data service loads the XML, keeps the first 500 records and counts how many records there are (slow).

As explained above, all these steps slow down processing the job, as the data service, the job engine and dacapo can all be running on different machines.

Therefore, we recommend that you call the CSV file through the URL instead. The following is an example of a URL:

file:/<PATH of csv file on the local drive>/example.csv