Social Science TeraGrid Gateway @ VirtualRDC

Lars Vilhuber and John Abowd
(Cornell U.)

Abstract

Tera-scale social science data are underutilized. Initially, serious confidentiality issues prevented most researchers from accessing these data. Significant research effort on projects that solve most of these confidentiality issues in combination with an expansion of the restricted-access model via Census Research Data Centers has begun to address this underutilization. Now that an increasing number of previously confidential data sources are finding their way into the public domain, the quantity of social science public-use data is once-again expanding dramatically.

The Social Science TeraGrid Gateway @ VirtualRDC proposes a method of unlocking those recently released data sources to allow much broader access by the social science research community. Most social science researchers face substantial hurdles when they wish to harness the power of large-scale computational clusters, in particular when using new, very large synthetic data sets with their unprecedented detail on people, jobs, and firms. The Social Science TeraGrid Gateway @ VirtualRDC will extend the VirtualRDC (http://www.vrdc.cornell.edu) model to allow support of tera-scale social science computing via the NSF-sponsored TeraGrid resources. The most widespread statistical software packages used by social scientists–SAS, Stata, and SPSS– are not available on the TeraGrid itself or on any of the servers at the borders of the TeraGrid with fast connections to it. When viewing the problem through the lens of the typical datadriven research process–extract, edit and transform data; transfer data to a computational location; and perform analysis–social science researchers are typically constrained in at least one of these steps when approaching the high performance computing clusters on the TeraGrid. For most data preparation, and for much analysis, the lack of standard statistical analysis and data preparation software packages is a serious impediment.

However, the typical social scientist's workstation or university-provided computational infrastructure does not have the resources to handle these very large data sets. Furthermore, the social scientist's workstation and the university-provided infrastructure do not have sufficiently fast data connectivity to transfer any large prepared data files to the TeraGrid for processing there.

The Social Science TeraGrid Gateway @ VirtualRDC aims to remedy bottlenecks in the first and second steps, with a focused expansion of resources at a critical location resulting in a highly useful gateway to the TeraGrid for the social sciences. The Social Science TeraGrid Gateway @ VirtualRDC (i) will allow researchers to perform the data preparation step using their "comfort-level" software packages, speeding up the data preparation phase, and (ii) do so on servers that have a fast connection to the TeraGrid, thus greatly speeding up the data-transfer process.