Interface, design, and computational considerations for D&R

Ryan Hafen

1. Hafen Consulting, LLC
*Contact author: rhafen@gmail.com

Keywords: Divide and Recombine, Large Complex Data, R, Hadoop

Divide and Recombine (D&R) is a statistical approach to deep analysis of large complex data [2]. In D&R, a large data set is broken into pieces in a meaningful way, statistical or visual methods are applied to each subset in an embarrassingly parallel fashion, and the results of these computations are recombined in a manner that yields a statistically valid result. D&R principles have been implemented in a software environment called Tessera (http://tessera.io) [1]. At the interface level, the Tessera analyst programs entirely in R, using two packages, datadr and trelliscope. The analyst uses datadr to read, manipulate, and specify D&R analytic operations on the data. The trelliscope package is used to visualize the divided data in a scalable and flexible manner [3, 4]. This interface ties to different computation and storage back ends in an agnostic and extensible manner. For very large data, Hadoop is used as a back end using the Rhipe [2] package to connect to the R Tessera interface. This talk covers some of the underlying goals, design, and interface requirements for programming with D&R and how these influenced the design of Tessera.

References


