Divide & Recombine with Tessera: Analyzing Larger and More Complex Data

William Cleveland

Statistics Department, Purdue
wsc@purdue.com

Keywords: Statistical Theory and Methods, *R*, **datadr**, **RHIPE**, Hadoop

Divide & Recombine (D&R) is a statistical approach to the analysis of large complex data. The data are divided into subsets by the analyst. Each of a collection of analytic methods is applied to the subsets by the analyst. The subset computations for each analytic method are independent, without communication among them. The outputs of each method can be recombined, or left as they are for further analysis. Statistical thinking, and statistical theory and methods research in D&R, hold much promise to make D&R a highly effective approach. Just as importantly, D&R leads to very simple parallel distributed computation. Tessera is an open source implementation of D&R. The front end is the *R* package **datadr**, which is a language for D&R. It makes programming D&R very simple and time efficient for the data analyst. At the back end is a distributed database and compute engine, so far most often Hadoop, which executes the *R* and **datadr** code of the analyst. Communication between **datadr** and Hadoop is provided by the *R* package **RHIPE**. D&R with Tessera can increase the size and complexity of the data that can be analyzed on a cluster, whether the cluster has small, moderate, or large hardware power. The data can have a memory size that is larger than the total cluster memory.