Tutorial:

Analysis and Visualization of Large Complex Data with Tessera

Ryan Hafen (Purdue University, Hafen Consulting, LLC)

Stephen F Elston (Quantia Analytics, LLC)

Background:

R is a powerful language for statistical analysis and visualization, with most of its power restricted to data of small or moderate size. Using Tessera, users readily visualize and analyze large complex data sets in a familiar R environment.

Developed over the past two years as part of the DARPA XDATA program, Tessera is an open source statistical computing environment. Tessera enables R users to perform deep analysis of large, complex data sets. Principal contributors to the project are statisticians and computer scientists at Purdue University and Pacific Northwest National Laboratory.

Tessera uses the Divide and Recombine (D&R) approach. In D&R, data is divided into meaningful subsets, embarrassingly parallel computations are performed on the subsets, and results are combined in a statistically valid manner. Using the R datadr package, Tessera provides a simple interface to distributed parallel back end computation environments such as Hadoop or Spark. Tessera includes a visualization component, Trelliscope, which provides a D&R approach for detailed, flexible, and interactive visualization of large complex data.

Tutorial overview:

R users will gain hands-on experience analyzing and visualizing data with Tessera. Using the famous ASA Airline data set, we will demonstrate what Tessera is and how to apply it. Attendees will develop a practical feel for using Tessera for statistical analysis and visualization.

The interactive tutorial examples are small enough to run on an attendee-provided laptop. The techniques learned can be quickly scaled up to a Tessera cluster for larger data sets.

Detailed Outline

This tutorial provides attendees hands-on experience using Tessera. We will cover the following topics in this tutorial:

Introduction to Tessera

- Divide and Recombine
- MapReduce
- Distributed data structures
- The datadr package
- The Trelliscope package
- System architecture options for Tessera
- Installation and configuration options for Tessera

Airline data analysis examples

- Distributed reading of raw text files
- Computing and visualizing summaries
- Conditioning variable division
- Dealing with complex data structures (e.g. multiple data sources)
- Analytic recombination examples
- Trelliscope display examples

Background Knowledge:

Attendees should have basic proficiency with R and RStudio.

Requirements for Interactive Session:

Attendees should have a laptop with the following installed:

- R 3.X
- A recent version of RStudio
- An up-to-date web browser, Chrome/Safari/Firefox
- The datadr package
- The trelliscope package

Potential Attendees:

This tutorial is intended for any R user who must analyze and understand large and complex data sets in a familiar environment.