Speaker: Nachum Shacham, eBay

Title: R in Big-Data Environments

Abstract:
Modern massively parallel processing (MPP) platforms like Hadoop and RDBMS, make Petabytes of structured and semi-structured records readily available for analysis. Running R on such platforms enables the processing of much larger datasets than is possible through the traditional single-server, RAM-based operation. We’ll describe our experience in using R to process big data on Teradata Enterprise Data Warehouse and on a large Hadoop cluster. We’ll review the challenges of running R in conjunction with these platforms and describe methods for accomplishing this task. The teradataR package that will be described, enables a PC-based R to use the warehouse resources to run statistical functions, e.g., regression, correlation, and parametric and nonparametric tests on warehouse-resident tables, while minimizing data transfer. We’ll also review our experience in using R for map-reduce processing of unstructured text files as well as tabular data on a 1000-node Hadoop cluster. Finally, we’ll review a case study, implemented in R, of comparative analysis of the cost of running queries on Teradata and Hadoop.