War stories from the data trenches

Richard A. Becker and Allan R. Wilks
AT&T Labs

Abstract

Data analysis for very large datasets is particularly challenging when the arrival of the data is relentless; it's like drinking from a fire hose. For several years we have been working with a 3 GB/day data stream, the transaction call detail of the AT&T long distance network. This talk will describe our experiences in taming the flow for several important applications, including fraud detection and database indexing. We will illustrate these experiences with anecdotes, and formulate general rules for dealing with huge amounts of data.