Column Databases Made Easy with R

Using mmap and indexing for high-performance data management and queries

Jeffrey A. Ryan
insight algorithmics, inc.
jeffrey.ryan@insightalgo.com
Chicago, Illinois USA

Keywords: column database, very large data, searching, queries, indexed searches

As data sets grow ever larger, the difficulty of accessing even subsets of data using R increases. By design, R stores all objects in memory and performs full table searches to extract relevant matches. Without an external database, users are limited to objects that must be a few times smaller than available memory. Often the only practical solution is to use an external database system interfaced with R.

The goal of the **indexing** package is to provide native R search semantics to very large data residing on-disk, organized by column, while reducing memory usage and dramatically speeding up search times. In effect, this alleviates the need to manage a separate external database. **indexing** accomplishes this by:

- Providing advanced index and search functionality to enable very fast and memory efficient boolean searches — including sparse, dense, and bitmap indexes. These tools work on most data objects in R, offering the ability to index vectors, matrix columns, entire data.frames, or atomic components of complex objects with nearly identical semantics. These can be used with resident memory objects, or in combination with disk-based data.
- Using optional memory-mapped files, data resides on disk until needed. Only relevant subsets are loaded into memory when needed for a search or extraction. This enables multi-gigabyte databases to be accessed as easily as if they resided in memory. It also allows for simple binary data files to be used, facilitating cross-application usage.
- Finally, and most critically, the design requires *no additional* R or database language skills, as it makes use of the standard R sub-setting and boolean search semantics.

The talk will examine the design, architecture, and implementation of the **indexing** and related **mmap** packages. Examples from a proprietary database of equity derivative option data covering over one million contracts, across sixty-seven million observations, and nineteen variables will be used to illustrate performance and functionality.

References

Jeffrey A. Ryan (2010). indexing and mmap packages
http://indexing.r-forge.r-project.org.