What's new

Major changes

- A new vectorized API supports a vectorized programming style, which can achieve C-like speeds when followed consistently. The old API is still the default and has undergone only cosmetic changes, but when efficiency is important and with small record sizes using the vectorized interface is recommended. On the other hand the old API allows simpler programs and has better compatibility with complex R objects used as keys and values. See the Introduction to the vectorized API for more information.
- In support of the vectorized API, there are completely new C implementations of serialization and deserialization from and to the typedbytes serialization format. This format is also important to exchange data with other members of the Hadoop system as typedbytes is the preferred serialization format for non-java streaming applications (HADOOP-1722). The implementation supports the most common and useful cases for R users but is not yet fully adherent to the specification yet.
- All formats are compatible with the new vectorized API but in particular “text” and “csv” readers and writers are substantially faster and the “csv” reader is even more stable when using the vectorized API.
- The support for structured data, while still evolving, has been extended to the map phase when using the vectorized API. that is we can have multiple records parsed at once and passed to the map function as a data frame. We plan to eventually have an “uninterrupted structured path” from input to output whereby the user would only manipulate data frames and costly conversions from and to data frames would be avoided. Issue #102 has been created to gather ideas and track progress on this.

Minor changes

- Updated whirr scripts run R 2.14, can optionally use lzo compression and work around unavaialbility of some packages from CRAN
- Packages are loaded on the cluster nodes with require to avoid failure when not available, nontheless your code will fail if such packages are needed by the map and reduce functions. The tradeoff here is that one doesn't have to detach packages before calling mapreduce when those packages are neither needed nor available on the nodes but the errors when packages are needed will be less intuitive (but a warning will be in stderr before the error, which should help)
- Fixed a bug (#111) whereby from.dfs would fail when tmp dir and destination are on two different volumes when the local backend is active (thanks Saar).