Robot Army:
A Distributed System for the Casual Manipulation of Massive Data Sets

Ira Woodhead
H5
71 Stevenson St.
San Francisco, CA 94105
ira@H5.com

ABSTRACT

The Robot Army is a command-oriented general purpose distributed processing system, allowing transform and aggregation operations over any size data set. It has been used successfully in a variety of e-Discovery tasks to produce high precision with high recall retrieval for collections of tens of millions of documents. The corpus format consists of ordered sets of untyped, variable length records of any form. Command orientation, as contrasted with code orientation, provides significant benefits. The user can leverage skills by invoking largely the same commands as on a single machine. One need not even write new programs at all if they already exist and fit the mapreduce model. For example, many standard unix utilities can be used as is. In addition, writing new programs is made simple by the fact that the system is by nature *language agnostic*. In contrast to Hadoop and Disco, there is no library API to learn, only STDIN and STDOUT; this allows the programmer to develop and test on a single machine with local sample data, then run on multiple machines with no customization necessary. In addition, Robot Army is small in lines of code (less than 1k loc), easy to understand and modify, and easy to install. We are planning to make it freely available as open source software.