SUMMARY

MAPREDUCE IS GOOD ENOUGH?

Large class of algorithms cannot use the concept of MapReduce implementation even though those problems can be easily implemented in MapReduce. Three large classes of problems server as poster children for MapReduce bashing.

Iterative Graph Algorithm

This algorithm PageRank defines the stationary distribution over vertices by a random walk over the graph, allowing random jumps to any other vertex in the graph. They have limitations of high startup costs with a lower bound on iteration time. Combiners and other local aggregation problems cannot fully solve the problem. The graph structure is structured wasted effort.

Gradient Descent

It solved the functional optimization problem. Each training data is processed in parallel and compute the partial contribution to the gradient which is emitted as an intermediate key-value pair and shuffled to a single reducer. Reducer sums up all gradient contributions and updates the model parameters. They have limitations of high startup costs, reducer must wait for all the mappers to finish, the combination of stragglers and using only a single reducer potentially causes poor cluster utilization.

Expectation Maximization

EM is an iterative algorithm that find a successive series of parameter estimates. EM iteration is typically implemented as a Hadoop job, set up the iterations and check for convergence. E-Step is performed in the mappers and M-step is performed in the reducers.

The Hadoop stack has already become the de facto general purpose, large scale data processing algorithm. Complete, end-to-end, large data solutions involve heterogeneous data sources and must integrate different types of processing, relational processing, graph analysis, text mining, machine learning. Map Reduce provides map and reduce which can be composed into more complex data flows. The data management and distributed systems have developed and refined a large bag of tricks over the several decades. Approach of incrementally refining Hadoop has a greater chance of making impact than a strategy of abandoning Hadoop. Open source everything and releasing of the software should be the default for any work. This represents a great potential for collaborations between academia and industry.