High Level GTA Programming Framework on MapReduce

We propose a high level programming framework for users to simply write GTA programs which can be deployed on large Hadoop cluster.

The Schematic Diagram

- Light-weight Programming interfaces
- Filter-embedding fusion
- Efficient MR programs

Users can choose necessary build-blocks provided by our framework

The specifications can be optimized to efficient list homomorphisms

Fully parallel and scalable Hadoop jobs are generated automatically

Programing using GTA - Building Blocks

- Even-Maximum-Prefix-Sum problem
- Knapsack problem
- Maximum Ascending Segments problem
- Count n-length subs problem

Generators
- Inits
- Subs
- Segs

Testers
- EvenSum
- LimitSum
- LimitLength
- Ascending

Aggregators
- MaxValue
- CountAll
- ...

Programming with GTA is just to define the specification by using GTA build-blocks

Evaluation on Hadoop Clusters

- Double the nodes of cluster can always gain 1.7 - 1.9 times speedup

Conclusions

- Performance: Good speedup and scalability
- Programmability: Clear and Light-weight interface
- Practicability: Various problems have been resolved

Reference