

# Real-time GDA

## ABSTRACT

### ACM Reference format:

. 2016. Real-time GDA. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17)*, 1 pages.  
DOI: 10.475/123\_4

## 1 MODEL

Let  $\mathcal{L}$  be the set of sites that holds data and runs tasks. For each inter-site WAN link, let  $B_{l_1}^{l_2}$  be the bandwidth from site  $l_1 \in \mathcal{L}$  to site  $l_2 \in \mathcal{L}$ . We assume that the bandwidths are stable within the time frame of doing real-time data analytics.

### 1.1 Single MapReduce Query

We perform the map tasks at the sites that contain the associated data and denote  $D_{l_1}$  as the output data from all of the map tasks at site  $l_1$ . The fraction of reduce tasks assigned to site  $l_2$  is denoted as  $r_{l_2}$  which is also the fraction of all other sites' data that must be transferred through the WAN to  $l_2$ . This means that the total WAN usage is for a given task distribution  $r$  is:

$$\sum_{l_1} \sum_{l_2 \neq l_1} D_{l_1} r_{l_2} \quad (1)$$

If we are given a start time  $s$  after the map steps are all completed and a data shuffle deadline  $t$  for the reduce tasks, then the completion time is bounded as such:

$$s + \max_{(l_1, l_2): l_2 \neq l_1} \left\{ \frac{D_{l_1} r_{l_2}}{B_{l_1}^{l_2}} \right\} \leq t \quad (2)$$

which is caused by the heterogeneous WAN bandwidth and has a bottlenecking link(s).

## 2 PROBLEM FORMULATIONS

### 2.1 Single MapReduce Query

*Minimize WAN usage.* When minimizing WAN usage for a MapReduce data shuffle we have the following optimization problem:

$$\min_r \sum_{l_1} \sum_{l_2 \neq l_1} D_{l_1} r_{l_2} \quad (3a)$$

$$\text{s.t.} \quad \sum_{l_2} r_{l_2} = 1 \quad (3b)$$

$$r_{l_2} \geq 0 \quad \forall l_2 \quad (3c)$$

*Feasibility to meet deadline.* We want to find a feasible  $r$  so that the data shuffle finishes at or before  $t$ :

$$\text{find } r \text{ s.t.} \quad \frac{D_{l_1} r_{l_2}}{B_{l_1}^{l_2}} \leq t - s \quad \forall (l_1, l_2) : l_2 \neq l_1 \quad (4a)$$

$$\sum_{l_2} r_{l_2} = 1 \quad (4b)$$

$$r_{l_2} \geq 0 \quad \forall l_2 \quad (4c)$$

*Minimize WAN usage given a shuffle deadline.* When minimizing WAN usage for a MapReduce data shuffle for a given deadline  $t$  we have the following optimization problem:

$$\min_r \sum_{l_1} \sum_{l_2 \neq l_1} D_{l_1} r_{l_2} \quad (5a)$$

$$\text{s.t.} \quad \frac{D_{l_1} r_{l_2}}{B_{l_1}^{l_2}} \leq t - s \quad \forall (l_1, l_2) : l_2 \neq l_1 \quad (5b)$$

$$\sum_{l_2} r_{l_2} = 1 \quad (5c)$$

$$r_{l_2} \geq 0 \quad \forall l_2 \quad (5d)$$

## APPENDIX

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Conference'17, Washington, DC, USA

© 2016 Copyright held by the owner/author(s). 123-4567-24-567/08/06...\$15.00

DOI: 10.475/123\_4