Title: Exploiting Coarse-grained Reused-based Opportunities in Big Data Multi-Query Optimization

Authors: Radhya Sahal, Mohamed H. Khafagy, Fatma A. Omara

PII: S1877-7503(17)30614-2
DOI: http://dx.doi.org/doi:10.1016/j.jocs.2017.05.023
Reference: JOCS 692

To appear in:

Received date: 1-9-2016
Revised date: 20-3-2017
Accepted date: 25-5-2017

Please cite this article as: Radhya Sahal, Mohamed H.Khafagy, Fatma A.Omara, Exploiting Coarse-grained Reused-based Opportunities in Big Data Multi-Query Optimization, Journal of Computational Sciencehttp://dx.doi.org/10.1016/j.jocs.2017.05.023

Exploiting Coarse-grained Reused-based Opportunities in Big Data Multi-Query Optimization

Radhya Sahal, Faculty of Computers and Information, Cairo University, Egypt, *radhya.sahal@grad.fci-cu.edu.eg*

Mohamed H. Khafagy, Faculty of Computers and Information, Fayoum University, Egypt, *mhk00@fayoum.edu.eg*

Fatma A. Omara, Faculty of Computers and Information, Cairo University, Egypt
f.omara@fci-cu.edu.eg

Highlights

- Big Data multi-query optimization system for real-world distributed applications running over Hadoop-like infrastructure.
- Exploit coarse-grained of reused-based opportunities regarding data size and non-uniform data distribution.
- Considering I/O speed of Hadoop storage such as Hard Disk Drives (HDDs).
- A partial reused-based multi-query optimizer to retrieve non-derived results of partial queries.

Abstract

Multi-query optimization in Big Data becomes a promising research direction due to the popularity of massive data analytical systems (e.g., MapReduce, Flink). The multi-query is translated into jobs. These jobs are routinely submitted with similar tasks to the underling Big Data analytical systems. These similar tasks are considered complicated and computation overhead. Therefore, there are some existing techniques that have been proposed for exploiting sharing tasks in Big Data multi-query optimization (e.g., MRShare and Relaxed MRShare). These techniques are heavily tailored relaxed optimizing factors of fine-grained reused-based opportunities. In accordance with Big Data multi-query optimization, the existing fine-grained techniques are only concerned with equal tuples size and uniform data distribution. These issues are not applicable to the real-world distributed applications which depend on coarse-grained reused-based opportunities, such as non-equal tuples size and non-uniform data distribution. These two issues receive more-attention in Big Data multi-query optimization, to minimize the data read from or written back to Big Data infrastructures (e.g., Hadoop). In this paper, Multi-Query Optimization using Tuple Size and Histogram (MOTH) system has been proposed to consider the granularity of the reused-based opportunities. The proposed MOTH system exploits the coarse-grained of the fully and partially reused-based opportunities among queries with considering non-equal tuples size and non-uniform data distribution to avoid repeated computations. According to the proposed MOTH system, a combined technique has been introduced for estimating the coarse-grained reused-based opportunities horizontally and vertically. The horizontal estimation of non-equal tuples size has been done by extracting metadata in column-level, while the vertical estimation of non-uniform data distribution is concerned with using pre-computed histogram in row-level. In addition, the MOTH system estimates the coarse-grained reused-based opportunities with considering slow storage (i.e., limited physical resources or fewer allocated virtualized resources) to produce the accurate estimation of the reused results costs. Then, a cost-based heuristic algorithm has been introduced to select the best reused-based opportunity and generate an efficient multi-query execution plan. Because the partial reused-based opportunities have been considered, extra computations are needed to retrieve the non-derived results. Also, a partial reused-based optimizer has been tailored and added to the proposed MOTH system to reformulate the generated multi-query plan to improve the shared partial queries. According to the experimental results of the proposed MOTH system using TPC-H benchmark, it is found that multi-query execution time has been reduced by considering the granularity of the reused results.

Keywords: *Big Data; Multi-Query Optimization; Coarse-grained; Sharing Opportunity; Reused-based;*