ملخص البحـث رقـــم (5)

عنوان البـــــحث باللغة الانجليزية:

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

اسماء المؤلفين:

RadhyaSahal, Mohammed H. Khafagy and Fatma A. Omara

مكان النشر و تاريخه:

Journal of Computational Science, Elsevier,2017

ملخص البـــــحث باللغة الانجليزية :

Multi-query optimization in Big Data becomes a promising research direction due to the popularity ofmassive data analytical systems (e.g., MapReduce and 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 someexisting techniques that have been proposed for exploiting sharing tasks in Big Data multi-query opti-mization (e.g., MRShare and Relaxed MRShare). These techniques are heavily tailored relaxed optimizingfactors 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 datadistribution. These issues are not applicable to the real-world distributed applications which dependon 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 minimizethe data read from or written back to Big Data infrastructures (e.g., Hadoop). In this paper, Multi-QueryOptimization 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 thefully and partially reused-based opportunities among queries with considering non-equal tuples size andnon-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 byextracting metadata in column-level, while the vertical estimation of non-uniform data distribution isconcerned with using pre-computed histogram in row-level. In addition, the MOTH system estimates thecoarse-grained reused-based opportunities with considering slow storage (i.e., limited physical resourcesor 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 opportunityand generate an efficient multi-query execution plan. Because the partial reused-based opportunitieshave been considered, extra computations are needed to retrieve the non-derived results. Also, a partialreused-based optimizer has been tailored and added to the proposed MOTH system to reformulate thegenerated multi-query plan to improve the shared partial queries. According to the experimental resultsof the proposed MOTH system using TPC-H benchmark, it is found that multi-query execution time hasbeen reduced by considering the granularity of the reused results

البحث مشتق من رسالة علمية

يقع البحث ضمن مجالات البحث بالقسم العلمي