Journal of Network and Computer Applications 34 (2011) 1035–1042

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Optimization of sub-query processing in distributed data integration systems

Gang Chen, Yongwei Wu*, Jia Liu, Guangwen Yang, Weimin Zheng

Department of Computer Science and Technology, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 28 October 2009
Received in revised form
21 April 2010
Accepted 7 June 2010
Available online 15 June 2010

Keywords:
Cloud computing
Grid computing
Data integration
Query
Data flow

ABSTRACT

Data integration system (DIS) is becoming paramount when Cloud/Grid applications need to integrate and analyze data from geographically distributed data sources. DIS gathers data from multiple remote sources, integrates and analyzes the data to obtain a query result. As Clouds/Grids are distributed over wide-area networks, communication cost usually dominates overall query response time. Therefore we can expect that query performance can be improved by minimizing communication cost.

In our method, DIS uses a data flow style query execution model. Each query plan is mapped to a group of μEngines, each of which is a program corresponding to a particular operator. Thus, multiple sub-queries from concurrent queries are able to share μEngines. We reconstruct these sub-queries to exploit overlapping data among them. As a result, all the sub-queries can obtain their results, and overall communication overhead can be reduced. Experimental results show that, when DIS runs a group of parameterized queries, our reconstructing algorithm can reduce the average query completion time by 32–48%; when DIS runs a group of non-parameterized queries, the average query completion time of queries can be reduced by 25–35%.

1. Introduction

As cloud and grid computing is becoming more and more popular, increasing number of applications needs to access and process data from multiple distributed sources. For example, a bioinformatics application needs to query autonomous databases across the world to access different types of proteins and protein–protein interaction information located at different storage clouds.

Data integration in Clouds/Grids is a promising solution for combining and analyzing data from different stores. Several projects (e.g., OGSA-DQP Lynden et al., 2009; CoDIMS-G Fontes et al., 2004; and GridDB-Lite Narayanan et al., 2003) have been developed to study data integration in distributed environments. For example, OGSA-DQP (Lynden et al., 2009) is a service-oriented, distributed query processor, which provides effective declarative support for service orchestration. It is based on an infrastructure consisting of distributed services for efficient evaluation of distributed queries over OGSA-DAI wrapped data sources and analysis resources available as services.

Queries to data integration systems are generally formulated in virtual schemas. Given a user query, a data integration system typically processes the query by translating it into a query plan and evaluating the query plan accordingly. A query plan consists of a set of sub-queries formulated over the data sources and operators specifying how to combine results of the sub-queries to answer the user query. As Clouds/Grids are generally built over wide-area networks, high communication cost is the main reason of leading to slow query response time. Therefore, query performance can be improved by minimizing communication cost. In this paper, our objective is to reduce communication overhead and therefore improve query performance, through optimizing sub-query processing.

We optimize sub-query processing by exploiting data sharing opportunities among sub-queries. IGNITE is a method proposed in Lee et al. (2007) to detect data sharing opportunities across concurrent distributed queries. By combining multiple similar data requests issued to the same data source, and further to a common data request, IGNITE can reduce communication overhead, thereby increase system throughput. However, IGNITE does not utilize parallel data transmission so that it does not always improve query performance. Our approach proposed here enhances IGNITE by addressing its drawbacks so that query performance in distributed systems can be further improved.

Our data integration system employs an operator-centric data flow execution model, also proposed in Harizopoulos et al. (2005). Each operator corresponds to a μEngine, which has local threads for data processing and data dispatching. Queries are processed by routing data through μEngines. All the μEngines work in parallel, thus they can fully utilize intra-query parallelism. Based

* Corresponding author. Tel.: +86 10 62796341.
E-mail addresses: c-g05@mails.tsinghua.edu.cn (G. Chen), wuyw@tsinghua.edu.cn (Y. Wu), liu-jia04@mails.tsinghua.edu.cn (J. Liu), ygw@tsinghua.edu.cn (G. Yang), zwm-dcs@tsinghua.edu.cn (W. Zheng).

1084-8045/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jnca.2010.06.007