Information Systems 37 (2012) 393–411

Contents lists available at ScienceDirect

Information Systems

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

OLAP query reformulation in peer-to-peer data warehousing

M. Golfarelli [a], F. Mandreoli [b,c], W. Penzo [a,c], S. Rizzi [a,*], E. Turricchia [a]

[a] DEIS - University of Bologna, V.le Risorgimento 2, 40136 Bologna, Italy
[b] DII - University of Modena and Reggio Emilia, Via Vignolese 905/b, Modena, Italy
[c] IEIIT - BO/CNR, V.le Risorgimento 2, 40136 Bologna, Italy

ARTICLE INFO

Available online 14 June 2011

Keywords:
Business intelligence
OLAP
Distributed databases
Query reformulation

ABSTRACT

Inter-business collaborative contexts prefigure a distributed scenario where companies organize and coordinate themselves to develop common and shared opportunities, but traditional business intelligence systems do not provide support to this end. To fill this gap, in this paper we envision a peer-to-peer data warehousing architecture based on a network of heterogeneous peers, each exposing query answering functionalities aimed at sharing business information. To enhance the decision making process, an OLAP query expressed on a peer needs to be properly reformulated on the local multidimensional schemata of the other peers. To this end, we present a language for the definition of mappings between the multidimensional schemata of peers and we introduce a query reformulation framework that relies on the translation of mappings, queries, and multidimensional schemata onto the relational level. Then, we formalize a query reformulation algorithm and prove two properties: correctness and closure, that are essential in a peer-to-peer setting. Finally, we discuss the main implementation issues related to the reformulation setting proposed, with specific reference to the case in which the local multidimensional engines hosted by peers use the standard MDX language.

1. Introduction

Business intelligence (BI) transformed the role of computer science in companies from a technology for passively storing data into a discipline for timely detecting key business factors and effectively solving strategic decisional problems. However, in the current changeable and unpredictable market scenarios, the needs of decision makers are rapidly evolving as well. To meet the new, more sophisticated user needs, a new generation of BI systems (often labeled as *BI* 2.0) has been emerging during the last few years.

One of the key features of BI 2.0 is the ability to become collaborative and extend the decision-making process beyond the boundaries of a single company [40]. Indeed, collaborative BI has been predicted to be the main BI trend for 2011 [30]. Users need to access information anywhere it can be found, by locating it through a semantic process and performing integration on the fly. This is particularly relevant in inter-business collaborative contexts where companies organize and coordinate themselves to share opportunities, respecting their own autonomy and heterogeneity but pursuing a common goal. In such a complex and distributed business scenario, traditional BI systems—that were born to support stand-alone decision making—are no longer sufficient to maximize the effectiveness of monitoring and decision making processes. Accessing local information is no more enough, users need to transparently and uniformly access information scattered across several heterogeneous BI platforms [25].

* Corresponding author.
E-mail addresses: matteo.golfarelli@unibo.it (M. Golfarelli), federica.mandreoli@unimo.it (F. Mandreoli), wilma.penzo@unibo.it (W. Penzo), stefano.rizzi@unibo.it (S. Rizzi), elisa.turricchia2@unibo.it (E. Turricchia).

0306-4379/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.is.2011.06.003

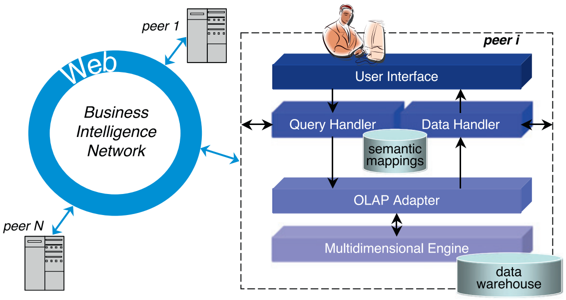

Fig. 1. Envisioned architecture for a BIN.

To fill this gap, we envision a peer-to-peer data warehousing architecture called *Business Intelligence Network* (BIN) and sketched in Fig. 1. A BIN is an architecture for sharing BI functionalities across a dynamic and collaborative network of heterogeneous and autonomous peers. Each peer is equipped with an independent data warehouse system, that relies on a local multidimensional schema to represent the peer's view of the business and exposes OLAP query answering functionalities (based for instance on the MDX language, a de-facto standard for querying multidimensional databases [36]) aimed at sharing business information, in order to enhance the decision making process and create new knowledge. The main benefits the BIN approach aims at delivering to the corporate world are the possibility of building new inter-organizational relationships and coordination approaches, and the ability to efficiently manage inter-company processes and safely sharing management information besides operational information [17].

The core idea of a BIN is that of enabling users to transparently access business information distributed over the network. A typical interaction sequence is the following:

1. A user formulates an OLAP query q by accessing the local multidimensional schema exposed by her peer, p.
2. Query q is processed locally on the data warehouse of p.
3. At the same time q is forwarded to the network.
4. Each involved peer locally processes the query on its data warehouse and returns its results to p.
5. The results are integrated and returned to the user.

The local multidimensional schemata of peers are typically heterogeneous; so, before a query issued on a peer can be forwarded to the network, it must be first *reformulated* according to the multidimensional schemata of the destination peers. In line with the approach adopted in *Peer Data Management Systems* (PDMSs) [24], query reformulation in a BIN is based on *semantic mappings* that mediate between the different multidimensional schemata exposed by two peers, i.e., they describe how the concepts in the multidimensional schema of one peer map onto those of another peer.

Direct mappings cannot be realistically defined for all the possible couples of peers. So, to enhance information sharing, a query q issued on p is forwarded to the network by first sending it to the neighborhood of p; then, each peer in this neighborhood in turn sends q to its neighborhood, and so on. In this way, q undergoes a chain of reformulations along the peers it reaches, and results are collected from any peer that is connected to p through a path of semantic mappings.

The approach outlined above is reflected by the internal architecture of each peer, sketched in the right side of Fig. 1, whose components are:

1. *User interface*. A web-based component that manages bidirectional interaction with users, who use it to visually formulate OLAP queries on the local multidimensional schema and explore query results.
2. *Query handler*. This component receives an OLAP query from either the user interface or a neighboring peer on the network, sends that query to the OLAP adapter to have it locally answered, reformulates it onto the neighborhood (using the available semantic mappings), and transmits it to the peers in that neighborhood.
3. *Data handler*. When the peer is processing a query that was locally formulated, the data handler collects query results from the OLAP adapter and from the peers, integrates them, and returns them to the user interface. When the peer is processing a query that was formulated on some other peer p, the data handler just collects local query results from the OLAP adapter and returns them to p.
4. *OLAP adapter*. This component adapts queries received from the query handler to the querying interface exposed by the local multidimensional engine.
5. *Multidimensional engine*. It manages the local data warehouse according to the multidimensional schema representing the peer's view of the business, and provides MDX-like query answering functionalities.

Interactions between peers are based on a message-passing protocol.

Query answering in a BIN architecture poses several research challenges, ranging from languages and models for semantic mediation to query reformulation issues and proper techniques and data structures for the query processing phase. Much work has been done on these issues in the context of PDMSs (e.g., [32–34]) and relational databases [38], however those results are not directly applicable in the OLAP scenario presented by the BIN.

In this paper we introduce the foundations for enabling query reformulation in a BIN. In particular, we focus on the definition of a language for semantic mapping specification to cope with the semantic heterogeneity of peers' schemata and on the specification of rules for reformulating OLAP queries. Reformulating an OLAP query is a challenging task due to the presence of aggregation and to the possibility of having information represented at different granularities and under different perspectives in each peer. The original contributions we give are:

- We present a language for the definition of semantic mappings between the schemata of peers, using predicates that are specifically tailored for the multidimensional model (Section 4). To overcome possible differences in data formats, mappings can be associated