Journal of Database Management
Volume 29 • Issue 2 • April-June 2018

An Efficient Stochastic Update Propagation Method in Data Warehousing

Bijoy Bordoloi, Southern Illinois University Edwardsville, Edwardsville, USA
Bhushan Kapoor, California State University - Fullerton, Fullerton, USA
Tim Jacks, Southern Illinois University Edwardsville, Edwardsville, USA

ABSTRACT

This article develops a stochastic update propagation method for an operational data store (ODS) in data warehousing (DW) environments where data storage (and retrieval) is required as a sum of data at distributed source nodes. The authors' proposed method results in less network traffic (as compared with the real-time method) due to update propagation required because of changes in source data. More importantly, the method allows system users to place limits on the discrepancy between the source data and the ODS data that could result due to a time lag between source data changes and the update operation. Finally, the pre-specified limits on the discrepancy are maintained while accounting for two crucial factors in distributed systems: 1) some nodes are situated on more congested network links, and 2) some of the links on the network are less reliable. Real-time data propagation does not account for these frequently encountered networking concerns.

KEYWORDS

Data Communications, Data Warehouse, ODS, Stochastic Processes

INTRODUCTION

The data warehouse (DW) continues to increase in importance as the core foundation of any Business Intelligence (BI) strategy. The DW and BI market reached $10.8 billion in 2011 and continues to be a top priority for CIOs (Demirkan & Delen, 2013). A data warehouse is a special type of centralized data storage facility in a distributed organizational information system which consolidates and integrates data from many different sources and presents it in an aggregate format to support decision making activities of middle or higher-level management personnel (Inmon & Hackathorn, 1994).

An Operational Data Store (ODS) is a crucial component of many DW architectures. It acts as an immediate staging area to store integrated data from different transaction systems prior to ETL (Extract, Transform and Load) processing on the centralized data warehouse (Sujitparapitaya et al., 2003). Data warehouses can be mission-critical enablers of organizational and inter-organizational strategic information systems such as Customer Relationship Management (CRM) (Cunningham et al, 2006). Other examples where a data warehouse can support the business strategy include Business Process Management and Supply Chain Management (Ariyachandra & Watson, 2010). The distributed

DOI: 10.4018/JDM.2018040102

nature of data warehousing architecture requires that any change in the source data at distributed locations in the network be propagated to the central DW via the ODS on a regular basis (Yang et al., 2011). The amount of traffic that is added to the network due to update propagation activities depends upon the propagation method used. Propagation can be accomplished either in real time or after a time lag which typically is the case with data warehousing (Doka et al., 2011; Inmon, 2000).

Though the contribution to the overall network traffic is likely to be less in the delayed batch mode, its usefulness is diminished by the fact that it can potentially result in a temporary and unknown amount of discrepancy between the warehouse data and the data at the source nodes. This discrepancy may not, however, be problematic provided its amount remains within pre-specified and known limits. While real-time processing is what the BI industry is moving towards due to increased requirements for organizational speed and agility, the infrastructure requirements for real-time information using data streams and in-memory processing can be prohibitively expensive for many organizations. Hence, it is beneficial to look for ways to optimize the traditional delayed mode of data delivery.

Most DW research tends to focus on optimizing server processing and storage once the data has already arrived in the DW (Cundius & Alt, 2013), but there seems to be a lack of research that accounts for network reliability and/or latency in the context of ODS and DW. Overall performance of a DW system can be impacted by overloaded nodes on the network that connect all the sources of DW data (Doka et al., 2011). Network reliability can be impacted by natural disasters such as the Great East Japan Earthquake of 2011. Network reliability can also be caused by intentional actions like a Denial-of-Service attack or unintentional events like faulty hardware, software, or configuration errors.

Network latency due to congestion on the Internet continues to be a problem. According to Cisco, the amount of data being transferred over the Internet (667 exabytes in 2013) is growing faster than the ability of the network infrastructure to carry that data (Demirkan & Delen, 2013). While newer networking technologies (like high-speed Metro Ethernet) can resolve many WAN congestion issues, high bandwidth circuits are not available everywhere. Furthermore, there are very large differences in network reliability levels in developing countries (Chandra et al., 2012). It cannot be assumed that every ODS or data warehouse has data sources with high-speed network capabilities.

Based on the stochastic paradigm of controlled imprecision (Rachev et al., 2008), this paper attempts to develop an approach to update propagation for data warehousing environments where ODS/DW data storage (and retrieval) is required as a sum of data at individual source nodes. Our procedure results in less network traffic (as compared with the real-time method) due to update propagation required because of changes in source data. More importantly, the method allows system users to control, within pre-established probabilistic limits, the discrepancy between the source data and the ODS/DW data that could result due to a time lag between source data changes and the update operation. Finally, the pre-specified limits on the discrepancy are maintained while accounting for two crucial factors in distributed systems: the fact that some nodes are situated on more congested network links and that some of the links on the network are less reliable than others.

In the rest of the paper, we present a literature review of relevant DW research. The next section shows the DW architecture to which our proposed method is applicable. Our proposed method for update propagation is then presented and illustrated through an example. Our method is based on controlled imprecision and the update trigger point for each source node in the network is central to our method. In the following section, the update trigger points based on the results of the mathematical derivations (described in the Appendix), are used to simulate a distributed system. The simulation results are used to verify whether overall discrepancy levels are within the pre-established limits. The final sections provide some concluding remarks.

LITERATURE REVIEW

The term ‘data warehouse’ was coined in the early 1990s as “…a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management’s decisions…” (Inmon, 1993). A