Chapter 11
Message-Based Approach to Master Data Synchronization among Autonomous Information Systems

Dongjin Yu
Hangzhou Dianzi University, China

ABSTRACT
The evolution of networks and large scale information systems has led to the rise of data sources that are distributed, heterogeneous, and autonomous. As a result, the management of Master Data becomes more complex and of uncertain quality. This paper presents a novel message-based approach to the synchronization of Master Data among multiple autonomous information systems. Different than traditional approaches based on database triggers, the author adopts the optimistic bidirectional strategy with the process of two synchronization phases. By means of data service buses, it propagates synchronized Master Data through messages being passed along star-like cascading routes. Moreover, this approach could resolve possible data conflicts automatically using predefined attribute confidences and deducible current value confidences respectively. Finally, this paper presents the real case about synchronizing datasets among four separate but related systems based on the author’s novel message-based approach.

INTRODUCTION
The evolution of networks and systems of data management has led to the rise of multiple wide scale information systems with data sources of very different kinds. Indeed, these sources can be distributed, heterogeneous and autonomous. Consequently, the lack of consistent and accurate reference information becomes the major issues confronting IT applications today. It is not uncommon to come across multiple sets of redundant and inconsistent data on customers and other items of primary focus in an organization (Piprani, 2009). Master Data Management (MDM) is an emerging discipline focusing on integrating data split in multiple systems by defining a master repository formatted as a data warehouse (Menet & Lamolle, 2009). Here, Master Data basically refer to the key business information which may include reference data about customers, products,
employees, materials, suppliers, etc. Master Data often turn out to be non-transactional in nature. In this regard, Master Data can support transactional processes and operations, and more often, relate with comprehensive analytics and reporting.

Data synchronization is an automated action to make the distributed Master Data be consistent with each other and up-to-date (Lee, Kim, & Choi, 2004). Through synchronization, the inconsistent, or stale Master Data, could be replaced with the right ones timely and automatically. Synchronization is different with backups, though they all involve data duplication. Backing up one disk to the other refers to make an exact copy of first disk onto to the other, thus preserving all the data on first one. By synchronization, however, the information from one dataset will be copied to the other, so each dataset will have the same information.

Data synchronization usually occurs between two or more autonomous information systems with locally administrated data stores. Under some complicated circumstances, synchronization would even involve tens of systems. Synchronization generally happens continually because distributed Master Data always keep changing. In order to synchronize data in a faster mode, only the shared information needs to be copied instead of duplicating the whole datasets. Since the data synchronization may involve large amount of data transferred under wide area networks, it however tolerates moderate time delay. In other words, data are never fully synchronized and synchronization just helps the distributed inconsistent data to be more consistent.

Synchronizing distributed Master Data is unavoidable under some circumstances. Case studies have revealed the significance of adoption of data synchronization for large organizations (Fuller, Sankar, & Raju, 2009; Zucker & Wang, 2009). For instance, the enterprise global data view could be only constructed with the data collected from multiple autonomous systems distributed inside the organization. Meanwhile, for the sake of business collaboration, different applications also need to be coupled with the underlying data swapping across system boundaries. In both cases, Master Data from different systems representing the same entity may be inconsistent because they are supervised under different domains and in different ways.

Current research issues related with data synchronization mainly focus on its performance, flexibility and methods of data conflict resolution. This paper presents the data synchronization framework, called MDSIF. MDSIF provides the coherent data snapshots for all related datasets by sending, receiving, and storing Master Data encapsulated in messages. Furthermore, MDSIF resolves data conflicts by setting Attribute Confidences and Current Value Confidences respectively. The Attribute Confidence, determined beforehand by data managers, indicates the reliability of all values of certain attribute created from certain dataset. Meanwhile, the Current Value Confidence, set automatically according to its own original Attribute Confidence, represents the present confidence of each shareable value. When the conflict is detected, the value with higher Current Value Confidence is allowed to overwrite that with lower one. Different with the traditional approaches such as triggers-based framework, MDSIF allows optimistic bidirectional synchronization, and is therefore more flexible, extendable and easier to be administrated.

The rest of the paper is organized in the following manner. The next section presents the architecture of synchronization framework based on messages, i.e., MDSIF. Following the overall introduction of conflict resolving strategy, the resolving method in the formal way, and its evaluation metrics are given. The successfully implemented case is then illustrated afterwards. After presenting related works, the paper finally provides concluding remarks and offers future research directions.