

Using Information Arbitration to Integrate and Interpret Data with Inconsistencies

Fan Lu
IBM Corporation

A common challenge in the Consumer Data Industry is how to integrate data with inconsistencies from multiple sources into a single consolidated view. For example, a credit bureau needs to integrate consumer credit history reported by multiple credit lending institutions. Over the past few years, a novel approach, information arbitration, has emerged to produce *full attribution, reversible, and sequence neutral integrations*.

Traditional approaches of information integration rely heavily on the *Merge and Purge* process. This approach generates a single record from two inconsistent ones by keeping one while purging the other, or by merging inconsistent attributes into one record. Despite the processing efficiency of this approach, its deterministic rules may introduce integration errors and these errors may be difficult to reverse. An example is the integration of two records, John Smith and John Smith Jr. with the same address and phone number. In many Merge and Purge integrations the two records would be consolidated into one. However this consolidation may not be warranted, because two original records can represent two separate individuals in some cases.

To avoid the integration errors of Merge and Purge, the Consumer Data Industry has also used another approach, *Tips and Leads*, to compile consumer data. This approach relies on the full capture of data from multiple sources without direct consolidation. A typical user inquiry of the compiled database may yield multiple results and give a user “tips and leads” for further application based resolution. This approach typically lacks the service characteristics required for transaction based inquiry processing.

On the other hand, information arbitration is a data centric, arbitrated integration, and it places less emphasis on the deterministic rules, such as those used by Merge and Purge. Central to this approach is the full maintenance of attribution of the ingested records, the extensive associations of these attributes, and the resolution of inquiries based on predetermined context. For example, the ingested data from many sources show John Smith currently living on 100 Main Street, and a newly reported record shows John Smith living on 200 Elm Street. The inquiry of John Smith’s address will yield 100 Main Street based on the most common view, but it can also produce 200 Elm Street based on the latest reported view. As more sources report 200 Elm Street as John Smith’s address, the most common view will evolve accordingly. Thus the integration by arbitrated data is a sequence neutral and fully reversible process. It is also significant to note that the arbitration approach generates extensive attribute associations and enhanced semantic understandings.

Information arbitration has the potential to become a general purpose approach for full attribution, reversible, and sequence neutral integrations. Its extensive attribute

association generates enhanced semantic understanding of the underlying data. To fully realize the potential of this integration approach, it will require additional research to find an optimized processing architecture for attribute associations and an arbitration context aware query language.