

Data Integration at Scale: From Relational Data Integration to Information Ecosystems

Michael L. Brodie

Verizon Communications Inc., USA

ABSTRACT

Our world is increasingly data-driven. The growth and value of data continue to exceed all predictions. Potential for business opportunity, economic growth, scientific discovery, and even national security lie hidden in the vast and growing data collections distributed across our Digital Universe. Harvesting the value of data requires finding, integrating, and analyzing data distributed across our Digital Universe. Failing to integrate data could be as mild as losing business through a lack of an integrated view of relevant business opportunities and threats or as serious as not anticipating the most serious terrorist attack on America through an inability to integrate relevant information distributed across 17 US intelligence agencies and 28 intelligence databases.

Due to data being distributed across both entities and data repositories, data integration has become perhaps the most important data operation but certainly the most costly accounting for up to 40% of data processing budgets. Data integration requirements are growing dramatically due to astounding growth in semantics (i.e., increasingly sophisticated application and data modeling requirements) and scale (i.e., web-scale data and transaction types, volumes, and distribution).

Since the mid-1980's the Relational Data Model has been the bedrock of business data processing providing a simple, elegant, and powerful basis for storing, finding, and integrating data based on simplifying assumptions. First, business data is naturally tabular; second, tabular business data follows relational principles; and third, assuming that there is a single source of truth, data views can be integrated into a single view under a global schema and that views of the same information are mutually consistent. To the extent that application data and operations fit naturally and meaningfully into the relational model and follow the three assumptions, Relational Database Management Systems (RDBMSs) provide the most efficient and cost effective data integration solution on the planet and will continue to do so through hardware and engineering advances.

As semantic and scale requirements of business data grew in the 1990's relational integration had to be extended ideally internal to the relational database engine, otherwise externally in data integration tools. This led to a plethora of data integration tools and supporting infrastructures with sales exceeding \$3.3 billion in 2009 with an expected annual grow rate of 8%. These integration solutions are typically less efficient than relational integration and considerably more costly to acquire and more complex to deploy, yet respond to genuine, if esoteric, data integration requirements.

While relational integration in RDBMSs augmented by integration tools meet the requirements of many large-scale applications, they are increasingly less applicable to a growing class of very-large-scale Information Ecosystems that can involve thousands of information systems and databases. Most Fortune 500 companies have ten or more major organizations – Corporate Management, Sales, Marketing, Engineering, Product Development, etc. - each requiring it's own view of the enterprise from it's unique perspective. Such views are often derived (i.e., integrated) from 5,000-10,000 databases. The Information Ecosystem is the information systems, databases, workflows, people, and infrastructure required to build, maintain, and dynamically update (i.e., integrate) the myriad organizational views. While the relational model and some combination of the three simplifying assumptions may apply to each component information system or database, the Information Ecosystem does not. While an individual information system may be designed to simplify the semantics of the real world application that it represents, the complexity of the real world or of the interactions between 1,000's of information systems cannot be as readily simplified. That is, relational integration and integration tools may apply to an individual or to a small number of information systems, but they are less applicable to large numbers of information systems such as occur in Information Ecosystems.

Relational technology is one of the most successful computing technologies in history based in part on the elegance, simplicity, and power of the Relational Data Model. It established a prototype for the next thirty years of database developments. Until the .com- and Web-fuelled expansion of our Digital Universe, the majority of data intensive applications were built on the bedrock of relational technology. Applications that did not fit or could not be forced into relational databases were built as extensions to relational technology or as entirely separate solutions. Both methods were not successful such as the rise and fall of object-oriented databases.

It now appears that the Relational Data Universe is less than 15% of the Digital Universe. Yet harvesting the value of data in our Digital Universe still requires finding, integrating, and analyzing data distributed across the Digital Universe. If relational integration applies to less than 15%, what data integration solutions are required for the other 85%?

The rapid expansion of our Digital Universe – both by amazing applications spawned by the Web and by phenomenal advances in hardware speeds and costs - led to a continuous stream of semantic and scale requirements that far outpaced data integration solutions and has spawned or been spawned by new technologies such as Web search, Collaborative Filtering, MapReduce, the Semantic Web, Ontologies, Cloud computing, and a dizzying array of others.

This talk surveys the evolution of data integration requirements and solutions from the Relational Data Universe to the Digital Universe. It examines data integration solutions in the Relational Data Universe starting with the relational data model and the three simplifying assumptions. It considers the major advances in data integration solutions in terms of semantics; database engine capabilities including data format, data architecture, and data governance; and the integration methods of integration tools. We consider a very-large-scale Information Ecosystem to illustrate limits of data integration solutions in the Relational Data Universe and conclude with review of directions towards data integration beyond the Relational Data Universe including new data models, a profound new direction from the Database Community; Information Ecosystems; Semantic Technologies emerging from the Semantic Web community; and Cloud computing where data integration may reside as Integration-as-a-Service or Information-as-a-Service (IaaS either way!).

KEYWORDS: Data Integration, Databases, Information Ecosystems,

INDEX TERMS: H Information Systems, H.2 Database Management, D.2 Software Engineering, I.2 Artificial Intelligence.