# The Power Behind the Throne: Information Integration in the Age of Data-Driven Discovery

Laura Haas
IBM Research - Almaden
lmhaas@us.ibm.com

## Abstract

Integrating data has always been a challenge. The information management community has made great progress in tackling this challenge, both on the theory and the practice. But in the last ten years, the world has changed dramatically. New platforms, devices and applications have made huge volumes of heterogeneous data available at speeds never contemplated before, while the quality of the available data has if anything degraded. Unstructured and semi-structured formats and no-sql data stores undercut the old reliable tools of schema, forcing applications to deal with data at the instance level. Deep expertise in the data and domain, in the tools and systems for integration and analysis, in mathematics, computer science, and business are needed to discover insights from data, but rarely are all of these skills found in a single individual or even team. Meanwhile, the availability of all these data has raised expectations for rapid breakthroughs in many sciences, for quick solutions to business problems, and for ever more sophisticated applications that combine and analyze information to solve our daily needs.

These expectations raise the bar for integration technology, while opening the door for it to play a broader role. Integration has always been a key player in handling data variety, for example, but now more than ever must deal with scale (in the number of types as well as in the volume and speed of data). While data cleansing has been one step of an integration pipeline, this technology must be leveraged throughout data integration, so that the integration process is better able to deal with the uncertainty in data, offering means to eliminate or reduce it, or, to elucidate it by linking important contextual information, such as provenance and usage. The complexity of today's data-driven challenges in fact suggests that the integration process should be context-aware, so that data sets may be combined differently depending on the proposed usage.

In the Accelerated Discovery Lab, we support data scientists working with a broad range of data as they try to find the insights to solve problems of business or societal importance. Clearly, integration is essential to insight. However, integration has to be across more than just datasets and schemas, and it has to be done more dynamically and flexibly than the standard tools allow. It is needed at multiple levels: (1) to build rich (but flexible) collections of diverse data, (2) to tightly bind individual data points into entities, allowing deeper explorations and (3) to bring together data and context to enable re-use by users with differing expertise. We think of the environment we are building as an integration hub for data, people and applications. It allows users to import, explore and create data and knowledge, inspired by the work of others, while it captures the patterns of decision-making and the provenance of decisions. I will describe the environment we are creating, the advances in the field that enable it, and the challenges that remain.


*SIGMOD'15,* May 31–June 4, 2015, Melbourne, Victoria, Australia.
ACM 978-1-4503-2758-9/15/05.
http://dx.doi.org/10.1145/2723372.2723373

**ACM Classification**
H.m [Miscellaneous] – Information Integration

**Keywords**
Information integration; big data; data analytics

**BIO**

Laura Haas is an IBM Fellow and Director of IBM Research's Accelerated Discovery Lab. She was Director of Computer Science at IBM's Almaden Research Center from 2005 to 2011, and had worldwide responsibility for IBM Research's exploratory computer science program from 2009 - 2013. From 2001-2005, she led the Information Integration Solutions architecture and development teams in IBM's Software Group. Previously, Dr. Haas was a research staff member and manager at Almaden. She is best known for her work on the Starburst query processor, from which DB2 LUW was developed, on Garlic, a system which allowed integration of heterogeneous data sources, and on Clio, the first semi-automatic tool for heterogeneous schema mapping. She has received several IBM awards including an IBM Corporate Award for information integration technology, and the Anita Borg Institute Technical Leadership Award. Dr. Haas was Vice President of the VLDB Endowment Board of Trustees from 2004-2009, and is a member of the National Academy of Engineering and the IBM Academy of Technology, an ACM Fellow, and Vice Chair of the board of the Computing Research Association.