Letter from the Special Issue Editor

Data integration is a long-standing problem over the past few decades. As more than 80% time of a data science project is spent on data integration, it becomes an indispensable part in data analysis. In recent few years, tremendous progress on data integration has been made from systems to algorithms. In this issue, we review the challenges of data integration, survey data integration systems (e.g., Tamr, BigGorilla, Trifacta, PyData), report recent progresses (e.g., explaining data integration, data discovery on open data, big data integration pipeline for product specifications, integrated querying of table data and S3 data, crowd-based entity resolution and human-in-the-loop rule learning), and discuss the future of this filed.

The first paper, "Data Integration: The Current Status and the Way Forward" by Michael Stonebraker and Ihab F. Ilyas, discusses scalable data integration challenges based on the experience at Tamr, and highlights future directions on building end-to-end scalable data integration systems, such as data discovery, human involvement, data exploration, and model reusability.

The second paper, "BigGorilla: An Open-Source Ecosystem for Data Preparation and Integration" by Alon Halevy et al, presents BIGGORILLA, an open-source resource for data preparation and integration. The paper describes four packages – an information extraction tool, a schema matching tool, and two entity matching tools.

The third paper, "Self-Service Data Preparation: Research to Practice" by Joseph M. Hellerstein et al, reviews self-service data preparation, which aims to enable the people who know the data best to prepare it. The paper discusses the key tasks in this problem and reviews the Trifacta system on how to handle these tasks.

The fourth paper, "Toward a System Building Agenda for Data Integration (and Data Science)" by Anhai Doan et al, advocates to build data integration systems by extending the PyData system and developing more Python packages to solve data integration problems. The paper provides an integrated agenda of research, system building, education, and outreach. The paper also describes ongoing work at Wisconsin.

The fifth paper, "Explaining Data Integration", by Laura Haas et al, reviews existing data integration systems with respect to their ability to derive explanations. The paper presents a new classification of data integration systems by their *explainability* and discusses the characteristics of systems within these classes. The authors also present a vision of the desired properties of future data integration systems with respect to explanations.

The sixth paper, "Making Open Data Transparent: Data Discovery on Open Data" by Renée J. Miller et al, discusses the problem of data discovery on open data, e.g., open government data. The paper considers three important data discovery problems: finding joinable tables, finding unionable tables, and creating an organization over a massive collection of tables.

The seventh paper, "Big Data Integration for Product Specifications" by Divesh Srivastava et al, presents an end-to-end big data integration pipeline for product specifications. The paper decomposes the problem into different tasks from source and data discovery, to extraction, data linkage, schema alignment and data fusion.

The eighth paper, "Integrated Querying of SQL database data and S3 data in Amazon Redshift" by Yannis Papakonstantinou et al, discusses query planning and processing aspects in Redshift, which provides integrated, in-place access to relational tables and S3 objects. The paper proposes techniques to optimize Redshift.

The ninth paper, "Robust Entity Resolution Using a CrowdOracle" by Divesh Srivastava et al, studies the problem of crowdsourcing-based entity resolution. The paper summarizes the CrowdOracle pipelines and describes a common framework consisting of simple operations.

The last paper, "Human-in-the-loop Rule Learning for Data Integration" by Ju Fan and Guoliang Li, proposes human-in-the-loop rule learning for effective data integration. The approach first generates a set of candidate rules, proposes a machine-based method to learn a confidence for each rule using generative adversarial networks, and devises a game-based crowdsourcing framework to refine the rules.

I would like to thank all the authors for their insightful contributions. I hope you enjoy reading the papers.

Guoliang Li
Tsinghua University
Beijing, China