An Introduction to Dynamic Data Quality Challenges

ALAN G. LABOUSEUR and CAROLYN C. MATHEUS, Marist College

CCS Concepts: • **Information systems** → **Database design and models**; **Relational database model**; **Graph-based database models**; *Information systems applications;*

Additional Key Words and Phrases: Dynamic data quality, internet of things, big data, relational systems, graph systems

ACM Reference Format:
Alan G. Labouseur and Carolyn C. Matheus. 2017. An introduction to dynamic data quality challenges. J. Data and Information Quality 8, 2, Article 6 (January 2017), 3 pages.
DOI: http://dx.doi.org/10.1145/2998575

1. INTRODUCTION

We live in an evolving world. As time passes, data changes in content and structure, and thus becomes dynamic. Data quality, therefore, also becomes dynamic because it is an aggregate characteristic of data itself. Thus, our evolving world and Internet of Things (IoT) presents renewed challenges in data quality. IoT data is teeming with multivendor and multiprovider applications, devices, microservices, and automated processes built on social media, public and private datasets, digitized records, sensor logs, web logs, and much more. From intelligent traffic systems to smart healthcare devices, modern enterprises are inundated with a daily deluge of dynamic big data.

The primary characteristics of big data are *volume*, *velocity*, and *variety* [Abadi et al. 2014]. Techniques for managing *volume* and *velocity* have been under development for decades. While some work has been done on *variety*, integrating and analyzing data from diverse sources and formats still presents challenges. For example, much of the big data deluge is structured and much of it is not. This single dimension of variety inherent in today's IoT clearly illustrates there is no "silver bullet" and *one size does not fit all* [Abadi et al. 2014; Stonebraker and Cetintemel 2005, 2015]. It is important to note there are many other dimensions of variety beyond structure. We must consider possibilities arising from analyzing data in a dizzying range of data types found in varying time frames of differing granularity from diverse sources in our evolving and streaming world. Structure is but one example illustrative of many more general challenges that we use in this article to introduce dynamic data quality.

2. CHALLENGES AND RESEARCH DIRECTIONS

Many challenges of dynamic data quality stem from the concept of *fitness for use*, a foundational idea in data quality research [Madnick et al. 2009]. Considering the variety of data found in our daily deluge and given the fact that one size does not fit

Authors' addresses: A. G. Labouseur and C. C. Matheus, Marist College, School of Computer Science and Mathematics, 3399 North Road, Poughkeepsie, NY 12601 USA; emails: {Alan.Labouseur, Carolyn.Matheus}@Marist.edu.

DOI: http://dx.doi.org/10.1145/2998575

all, we must be prepared to alter the underlying representation of our data, casting it from one form to another, to ensure fitness for a given use at a given time. Many challenges present themselves when we are faced with preserving *fitness for use*, including determining what happens to data quality dimensions as we change the underlying characteristics, what data quality trade-offs occur when we cast data from one representation to another, whether or not we can enhance data quality as a side effect of changing its structure, and how to measure dynamic data quality in an environment of evolving data. With these specifics in mind, we present a general research challenge:

> What happens to data quality when casting underlying data storage from one structure to another?

Exploring the vast implications and plentiful research opportunities stemming from this general challenge is beyond the scope of this short article. Therefore, to illustrate the idea and provide a concrete starting point for additional research, we narrow our focus to a few challenges arising in relational- and graph-structured storage while posing data quality challenges across three dimensions. There are more data management domains to consider: document stores, key-value stores, object stores, array engines, and hybrids. There are also more data quality dimensions to consider in combination with those domains. This is just a start.

2.1. Relational and Graph Systems

Relational systems are well suited for managing data structured as tables of rows and columns, and performing common analytic tasks that graph systems are bad at, such as creating segmentations based on attributes and combining data based on matching values. Graph systems are well suited for managing data structured as vertices and edges, and performing common analytic tasks that relational systems are bad at, such as finding clusters, determining shortest paths, and computing influence. Valuable tool chains must provide analytics regardless of the form of the data. It is, therefore, necessary to take relational data and cast it as a graph, to take graph data and cast it as relations, and to understand the effects of these transformations on data quality.

2.2. Data Quality Dimensions

Numerous dimensions of data quality have been identified [Pipino et al. 2002; Wand and Wang 1996]. For the sake of brevity, we discuss three of these dimensions here.

Accessibility refers to the extent to which data are available and easily retrievable. Batini et al. [2009] propose using response or delivery time as a metric for accessibility. They further suggest that this measure is applicable to two types of data variety: (1) structured or "sensed" data such as digitized records, sensor logs, or web logs; and (2) unstructured data found in social media, public and private data sets, and data retrieved from the web. Using query performance as a proxy for response or delivery time to measure *accessibility*, one challenge lies in predicting trade-offs in query performance and how they might vary with dynamic workload. Along with research on estimating query performance in relational systems [Yin et al. 2015], recent work on predicting query performance in a distributed graph system [Labouseur et al. 2015b] may prove helpful in addressing these challenges.

Ease of manipulation refers to the extent to which data are simple to manipulate and apply to different tasks, such as how easily data can be updated and aggregated. While data may be in a form that is optimized for certain types of analyses (e.g., finding multi-hop influencers in a graph), that same form may prevent easy update or aggregation (e.g., changing common attributes in many vertices of a graph). Casting data to another form may ease update and aggregation while simultaneously making analysis more difficult. Pipino et al. [2002] propose that *ease of manipulation* can be

assessed by measuring the ratio of desired outcomes to total outcomes. One challenge is in determining which form of data is best based on how usage of the data evolves over time and how that influences the ratio of desired outcomes to total outcomes. We speculate that optimization and estimation systems such as Macrobase [Bailis 2015] could help preserve or predict various aspects of performance-related data quality.

Representation refers to the extent to which data are compactly represented, well organized, and well formatted. *Consistent representation* refers to the extent to which data are presented in a compatible form over time. As with *ease of manipulation*, a ratio measuring the desired outcomes as a portion of total outcomes can be used to gauge the representation of data. We speculate that polystores like BigDAWG [Duggan et al. 2015] and hybrid systems like Myria [Halperin et al. 2014] might be helpful in addressing these challenges in static relational and graph systems. For *consistent representation* over time we look to the long-explored subject of temporal relational databases and the new area of dynamic graph systems [Labouseur et al. 2015a].

3. CONCLUSION

The principle that one size does not fit all necessitates diverse strategies in building systems capable of analyzing and managing the IoT data deluge. Regardless of your approach to maintaining fitness for use—or any of the plethora of other data quality dimensions—the many challenges of dynamic data and dynamic data quality promise to provide research opportunities for a long time to come.

REFERENCES

Daniel Abadi, Rakesh Agrawal, Anastasia Ailamaki, Magdalena Balazinska, Philip A. Bernstein, Michael J. Carey, Surajit Chaudhuri, Jeffrey Dean, AnHai Doan, Michael J. Franklin, and others. 2014. The Beckman report on database research. *SIGMOD Records* 43, 3 (Dec. 2014), 61–70.

Peter Bailis. 2015. Marcobase. Retrieved from https://github.com/stanford-futuredata/macrobase.

Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* 41, 3, Article 16 (July 2009), 52 pages.

Jennie Duggan, Aaron J. Elmore, Michael Stonebraker, Magda Balazinska, Bill Howe, Jeremy Kepner, Sam Madden, David Maier, Tim Mattson, and Stan Zdonik. 2015. The BigDAWG polystore system. *SIGMOD Rec.* 44, 2 (Aug. 2015), 11–16.

Daniel Halperin, Victor Teixeira de Almeida, Lee Lee Choo, Shumo Chu, Paraschos Koutris, Dominik Moritz, Jennifer Ortiz, Vaspol Ruamviboonsuk, Jingjing Wang, Andrew Whitaker, Shengliang Xu, Magdalena Balazinska, Bill Howe, and Dan Suciu. 2014. Demonstration of the Myria big data management service. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. 881–884.

Alan G. Labouseur, Jeremy Birnbaum, Paul W. Olsen, Sean R. Spillane, Jayadevan Vijayan, Jeong-Hyon Hwang, and Wook-Shin Han. 2015a. The G* graph database: Efficiently managing large distributed dynamic graphs. *Distrib. Parallel Database* 33, 4 (2015), 479–514.

Alan G. Labouseur, Justin Svegliato, and Jeong-Hyon Hwang. 2015b. Distributed graph snapshot placement and query performance in a data center environment. In *Proceedings of 2015 International Conference on Computational Science and Computational Intelligence*. ACSE, IEEE CPS, 348–351.

Stuart E. Madnick, Richard Y. Wang, Yang W. Lee, and Hongwei Zhu. 2009. Overview and framework for data and information quality research. *J. Data Inf. Quality* 1, 1 (June 2009).

Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data quality assessment. *Commun. ACM* 45, 4 (April 2002), 211–218.

Michael Stonebraker and Ugur Cetintemel. 2005. "One size fits all": An idea whose time has come and gone. In *Proceedings of the 21st International Conference on Data Engineering (ICDE'05)*. 2–11.

Michael Stonebraker and Ugur Cetintemel. 2015. 10-Year Most Influential Paper Award Talk: One Size Fits All; 10 Years Later - One Size Fits None!!. From ICDE 2015. Remarks about their 10-year "test of time" award. Accessed: 2015-12-30. http://kdb.snu.ac.kr/data/stonebraker_talk.mp4.

Yair Wand and Richard Y. Wang. 1996. Anchoring data quality dimensions in ontological foundations. *Commun. ACM* 39, 11 (Nov. 1996), 86–95.

Shaoyi Yin, Abdelkader Hameurlain, and Franck Morvan. 2015. Robust query optimization methods with respect to estimation errors: A survey. *SIGMOD Rec.* 44, 3 (Dec. 2015), 25–36.

Received January 2016; revised September 2016; accepted September 2016