The Data Repurposing Challenge: New Pressures from Data Analytics

PHILIP WOODALL, University of Cambridge

Categories and Subject Descriptors: E.0 [**Data**]: General; H.0 [**Information Systems**]: General

General Terms: Data quality, Information quality, Data analytics, Data repurposing, Data reuse

Additional Key Words and Phrases: Business intelligence, Information repurposing, Information reuse

ACM Reference Format:

...

When data is collected for the first time, the data collector has in mind the data quality requirements that must be satisfied before it can be used successfully—i.e. the data collector ensures "fitness for use"—the commonly agreed upon definition of data quality [Wang and Strong 1996]. However, data that is repurposed [Woodall and Wainman 2015], as opposed to reused, must be managed with multiple different fitness for use requirements in mind, which complicates any data quality enhancements [Ballou and Pazer 1985]. While other work has considered context in relation to data quality requirements, including the need to meet multiple fitness for use requirements [Watts et al. 2009; Bertossi et al. 2011], in the current fast-paced environment of data repurposing for analytics and business intelligence, there are new challenges for dealing with multiple fitness for use requirements in the context of:

1. Ephemeral data use
2. Self-service data collection

Ephemeral data use is when the use of the data is either short-lived or, after collection and transformation, does not prove to be useful at all. The former situation occurs when data analytics results are only used, for example, once or twice (rather than regularly for day-to-day operational decisions). The latter situation occurs when users are performing data analytics with the aim of revealing business insights, which may never materialise because of the exploratory nature of the task. The problem is that discovering that the data is not useful occurs after effort has been expended extracting and transforming the data to get it to the point where it can be analysed. In this case, it is necessary to minimise the time and effort in transforming the data to satisfy the new fitness for use requirements. Otherwise, important results may be missed because of analysts/developers not being willing to invest the time it takes to transform the data to be fit for the new use. If one knows that in many cases the results will be discarded, then there is little incentive to invest large amounts of time and resources to transform the data each time. A specific example of this is investigating whether online public data, such as social media posts and news events, can be used with internal procurement data to provide advanced warning of parts supply shortages to a manufacturing organisation. Even in the exploratory case, effort is required to wrangle the data into a usable form, filter events that do not relate to the manufacturer's suppliers, link events to the relevant part deliveries etc. before any analysis can be done to determine its predictive capability. This effort is wasted for the organisation if it turns out that the data cannot provide useful predictions.

Self-service data collection is when users, rather than IT departments, extract, transform and produce their own reports from data [Schlesinger and Rahman 2015]. It emerged from both the increased pressure to perform data analytics for business intelligence and the fact that the IT department cannot always meet the increased data demands of the users. The problem is that if users are left to collect and transform the data themselves, how can an organisation be sure that they have the expertise to judge whether it has been done correctly and hence does actually meet the fitness for use requirements? Furthermore, in order to reliably use the data, the analysts must have good visibility of the assumptions and key facts of the data collection and prior transformations, otherwise there is a danger that they may be overlooked [Kennedy et al. 2015] leading to the drawing of inaccurate conclusions. This was the case in the example given by Veaux and Hand [2005] where a data analyst expected that the data they were using was from a direct measurement when, in fact, it was from the output of a simulation/model. Another example relates to a data analyst at a manufacturer who used expected delivery date and the goods storage date to identify poorly performing suppliers [Woodall and Wainman 2015]. Comparing these values made many suppliers appear to deliver late because the storage personnel would often wait until the following day before either storing the goods and/or recording them as being stored. These data fields were perfectly accurate for their primary purpose of inventory recording, but were not suitable for their repurposed use in calculations to identify poor supplier performance.

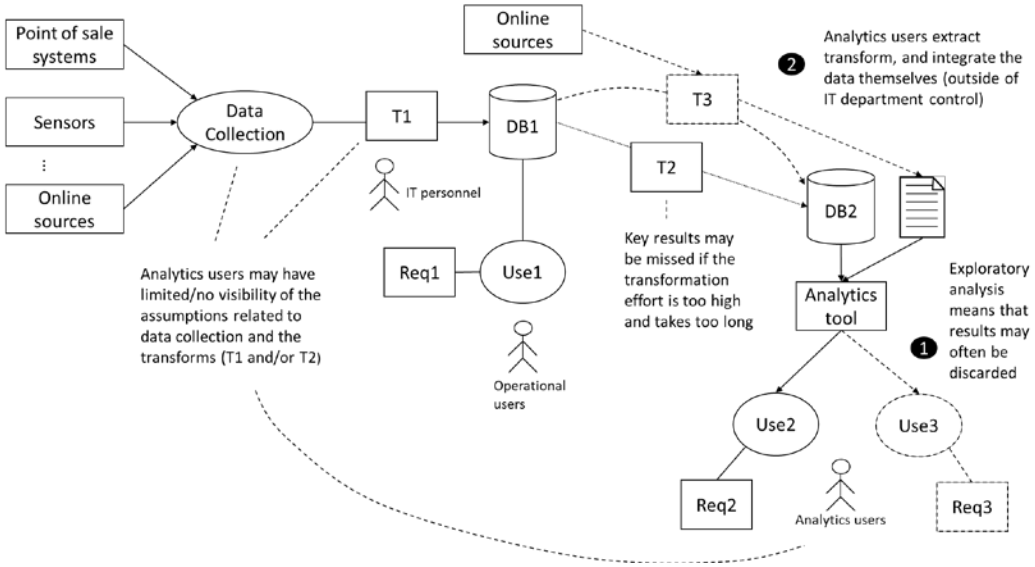

Figure 1: The two key pressures on data repurposing

These two pressures are illustrated in Figure 1, which shows how the data is collected, is transformed (T1), is being managed by the Information Technology (IT) personnel (in BD1) for its primary use (Use1), and according to data quality requirements (Req1). Analytics users may also take a copy of this data and collect new data from external sources, integrate and transform it (using T2, or T3 by themselves), hold a copy (in DB2 or their own files, such as spreadsheets), before running it through analytics tools for Use2 and Use3 (the results of which may be discarded immediately or within a short time-scale).

To address these challenges, developing new and enhancing existing data quality tools/methods which focus on further reducing both the time and effort to bring data to the analysis stage is needed. That is, methods in the prior stages of the data analysis pipeline: data acquisition, extraction, cleaning, integration, aggregation, and representation [Jagadish et al. 2014]. These could include, (semi) automated methods to discover and evaluate which data sources—including user-generated content [Lukyanenko et al. 2014]—contain data that is, or can be made, fit for purpose. For the stages after extraction, data quality aware transformation platforms could be designed to support self-service users to make rapid and potentially automated changes to data to enable it to quickly meet different fitness for use requirements. Predictions (e.g. using machine learning) of how data will need to be used/analysed in the future could enable "pre-transforming" or "pre-selection" of data so that it is ready for analysis when (or even before) the analyst needs it. These approaches must also capture key metadata, such as provenance and the context of the data before if it is cleaned, integrated, or aggregated into different forms. Using this metadata, fitness for use validation alerts could warn users when data is being used that does not meet the new quality requirements.

Advances in how intuitive these approaches are for non-expert users (who, unlike developers, may have no understanding of database structures and data profiling terms etc.) is needed throughout the data analysis pipeline. Finally, education of the importance of, and existing techniques in, data quality for data scientists is essential, now that they are starting to collect, transform, analyse and report on organisational data themselves outside of the control of information technology professionals.

REFERENCES

D. Ballou and H. Pazer. 1985. Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Manag. Sci.* 31, 2 (1985), 150–162.

L. Bertossi, F. Rizzolo, and L. Jiang. 2011. Data Quality Is Context Dependent. In M. Castellanos, U. Dayal, & V. Markl, eds. *Enabling Real-Time Business Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 52–67.

H.V. Jagadish et al. 2014. Big data and its technical challenges. *Commun. ACM* 57, 7 (July 2014), 86–94. DOI:https://doi.org/10.1145/2611567

O. Kennedy, Y. Yang, J. Chomicki, R. Fehling, Z.H. Liu, and D. Gawlick. 2015. Detecting the Temporal Context of Queries. In M. Castellanos, U. Dayal, T. B. Pedersen, & N. Tatbul, eds. *Enabling Real-Time Business Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 97–113.

R. Lukyanenko, J. Parsons, and Y.F. Wiersma. 2014. The IQ of the Crowd: Understanding and Improving Information Quality in Structured User-Generated Content. *Inf. Syst. Res.* 25, 4 (December 2014), 669–689. DOI:https://doi.org/10.1287/isre.2014.0537

P.A. Schlesinger and N. Rahman. 2015. Self-Service Business Intelligence resulting in disruptive technology. *J. Comput. Inf. Syst.* 56, 1 (2015), 11–21.

R.D. De Veaux and D.J. Hand. 2005. How to Lie with Bad Data. *Stat. Sci.* 20, 3 (August 2005), 231–238. DOI:https://doi.org/10.1214/088342305000000269

R.Y. Wang and D.M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst.* 12, 4 (1996), 5–34.

S. Watts, G. Shankaranarayanan, and A. Even. 2009. Data quality assessment in context: A cognitive perspective. *Decis. Support Syst.* 48, 1 (2009), 202–211.

P. Woodall and A. Wainman. 2015. Data Quality in Analytics: Key Problems Arising from the Repurposing of Manufacturing Data. In *International Conference on Information Quality (ICIQ)*. Cambridge, MA., 174–184.