DATA QUALITY IN BIG DATA ENVIRONMENTS

Marcel Altendeitering
„Garbage in, Garbage out“? – Why data quality matters

What a lack of data quality looks like…

- Research on two data sets from the domains of stock markets and airline flights, usually considered as highly reliable, showed that 70% of data items were inconsistent and 50% of them had ambiguous values [1].

- According to Gartner [3] poor data quality is responsible for an average of $15 million in costs per year and company. HBR [4] sums that up to $3 Trillion per year for the US economy.

Improved data quality leads to…

- Improved data-driven decision making [5]
- Better performance of a company [2]
What makes big data quality so difficult?

- Typical data quality issues (e.g. inconsistent, incomplete or duplicate values) are more difficult in big data environments due to the four V’s [6].
- Generally data quality issues increase proportionally to volume and variety, however some errors can increase exponentially [7].
- Traditional solutions like Data Warehouses are not sufficient to deal with data quality in big data [8].
Tackling big data quality by combining organizational and technical actions

- A combination of organizational (e.g. data governance rules) and technical (e.g. automated data quality control) actions is required [4, 9, 10]
- Research on solutions for big data quality is accelerating based on literature

Own illustration

# of solutions for big data quality in literature

Data lake architecture proposed by [9]
Marcel Altendeitering
Wissenschaftlicher Mitarbeiter
Emil-Figge-Straße 91, 44227 Dortmund
Telefon: +49 (0) 231 / 9 76 77-461
marcel.altendeitering@isst.fraunhofer.de
Sources


