HPI Kolloquium
28.04.2016, 16:00 Uhr

Hasso-Plattner-Institut, Vorlesungsgebäude, Auditorium 1
Campus Griebnitzsee, 14482 Potsdam

“Data Cleaning from Theory to Practice”

Prof. Ihab Ilyas
Cheriton School of Computer Science, University of Waterloo

Abstract

With decades of research on the various aspects of data cleaning, multiple technical challenges have been tackled and interesting results have been published in many research papers. Example quality problems include missing values, functional dependency violations and duplicate records. Unfortunately, very little success can be claimed in adopting any of these results in practice. Businesses and enterprises are building silos of home-grown data curation solutions under various names, often referred to as ETL layers in the business intelligence stack. The impedance mismatch between the challenges faced in industry and the challenges tackled in research papers explain to a large extent the growing gap between the two worlds. In this talk I claim that being pragmatic in developing data cleaning solution does not necessarily mean being unprincipled or ad-hoc. I discuss a subset of these practical challenges including data ownership, human involvement, and holistic data quality concerns. These new set of challenges often hinder current research proposals from being adopted in the real world. I also go through a quick overview of the approach we use in Tamr (a data curation startup) to tackle these challenges.

Short CV

Ihab Ilyas is a professor in the Cheriton School of Computer Science at the University of Waterloo. He received his PhD in computer science from Purdue University, West Lafayette. His main research is in the area of database systems, with special interest in data quality, managing uncertain data, rank-aware query processing, and information extraction. Ihab is a recipient of the Ontario Early Researcher Award (2009), a Cheriton Faculty Fellowship (2013), an NSERC Discovery Accelerator Award (2014), and a Google Faculty Award (2014), and he is an ACM Distinguished Scientist. Ihab is a co-founder of Tamr, a startup focusing on large-scale data integration and cleaning. He serves on the VLDB Board of Trustees, and he is an associate editor of the ACM Transactions of Database Systems (TODS).

Host: Prof. Dr. Felix Naumann