The Case for Small Data Management

Jens Dittrich, jens.dittrich@cs.uni-saarland.de

1 Abstract

Exabytes of data; several hundred thousand TPC-C transactions per second on a single computing core; scale-up to hundreds of cores and a dozen Terabytes of main memory; scale-out to thousands of nodes with close to Petabyte-sized main memories; and massively parallel query processing are a reality in data management. But, hold on a second: for how many users exactly? How many users do you know that really have to handle these kinds of massive datasets and extreme query workloads? On the other hand: how many users do you know that are fighting to handle relatively small datasets, say in the range of a few thousand to a few million rows per table? How come some of the most popular open source DBMS have hopelessly outdated optimizers producing inefficient query plans? How come people don't care and love it anyway? Could it be that most of the world's data management problems are actually quite small? How can we increase the impact of database research in areas when datasets are small? What are the typical problems? What does this mean for database research? We discuss research challenges, directions, and a concrete technical solution coined PDbF: Portable Database Files. This is an extended version of an abstract and Gong Show talk presented at CIDR 2015 (http://www.cidrdb.org/cidr2015/Papers/11_Abstract17DJ.pdf).

2 Biography

Jens Dittrich is a Full Professor of Computer Science in the area of Databases, Data Management, and Big Data at Saarland University, Germany. Previous affiliations include U Marburg, SAP AG, and ETH Zurich. He is also associated to CISPA (Center for IT-Security, Privacy and Accountability). He received an Outrageous Ideas and Vision Paper Award at CIDR 2011, a BMBF VIP Grant, a best paper award at VLDB 2014, two CS teaching awards in 2011 and 2013, as well as several presentation awards including a qualification for the interdisciplinary German science slam finals in 2012 and three presentation awards at CIDR (2011, 2013, and 2015).

His research focuses on fast access to big data including in particular: data analytics on large datasets, Hadoop MapReduce, main-memory databases, and database indexing. He has been a PC member and/or area chair of prestigious international database conferences such as PVLDB, SIGMOD, and ICDE. Since 2013 he has been teaching his classes on data management as flipped classrooms. See http://datenbankenlernen.de or http://youtube.com/jensdit for a list of freely available videos on database technology in German and English (about 80 videos in German and 80 in English so far).