Breaking the Chains: On Declarative Analysis and Independence in the Big Data Era

Volker Markl

Database Systems and Information Management
Technische Universität Berlin, Germany
volker.markl@tu-berlin.de

Abstract: Data management research, systems, and technologies have drastically improved the availability of data analysis capabilities, particularly for non-experts, due in part to low-entry barriers and reduced ownership costs (e.g., for data management infrastructures and applications). Major reasons for the widespread success of database systems and today's multi-billion dollar data management market include data independence, separating physical representation and storage from the actual information, and declarative languages, separating the program specification from its intended execution environment. In contrast, today's big data solutions do not offer data independence and declarative specification. As a result, big data technologies are mostly employed in newly-established companies with IT-savvy employees or in large well-established companies with big IT departments. We argue that current big data solutions will continue to fall short of widespread adoption, due to usability problems, despite the fact that in-situ data analytics technologies achieve a good degree of schema independence. In particular, we consider the lack of a declarative specification to be a major roadblock, contributing to the scarcity in available data scientists available and limiting the application of big data to the IT-savvy industries. In particular, data scientists currently have to spend a lot of time on tuning their data analysis programs for specific data characteristics and a specific execution environment. We believe that the research community needs to bring the powerful concepts of declarative specification to current data analysis systems, in order to achieve the broad big data technology adoption and effectively deliver the promise that novel big data technologies offer. In addition, we will present the vision of the Berlin Big Data Center (BBDC) with respect to combining machine learning and data management. A 4-page paper extended abstract of this presentation has appeared at PVLDB Volume 7.

Volker Markl is a Full Professor and Chair of the Database Systems and Information Management (DIMA) group at the Technische Universitat Berlin (TU Berlin) as well as an adjunct status-only professor at the University of Toronto. Earlier in his career, Dr. Markl lead a research group at FORWISS, the Bavarian Research Center for Knowledge-based Systems in Munich, Germany, and was a Research Staff member & Project Leader at the IBM Almaden Research Center in San Jose, California, USA. Dr. Markl has published numerous research papers on indexing, query optimization, lightweight information integration, and scalable data processing. He holds 7 patents, has transferred technology into several commercial products, and advises several companies and startups. He has been

speaker and principal investigator of the Stratosphere research project that resulted in the "Apache Flink" big data analytics system. Dr. Markl currently serves as the secretary of the VLDB Endowment and was recently elected as one of Germany's "digital minds" (Digitale Köpfe) by the German Informatics Society (GI).