Database Technology for Supporting the Analysis of Complex Monitoring Scenarios

Dr. Andreas Behrend

Abstract:

Situation awareness refers to the capability of systems to perceive an existing or predicted context that determines the values of variables in a changing environment. Today we realize that many applications have to deal with situation awareness: Customer Relationship Management, Supply Chain Management, patient care, power grid and cloud services management, as well as any IoT related application; just to name a few. Very often, database systems play a central role in these applications but despite the enhanced support for managing timestamped data, current database systems still lack mechanisms for handling temporal situations in which data may change frequently.

In this talk we present first results from an ongoing research project investigating these missing database features. In particular, we identify (i) the requirements for representing complex situations as spatio-temporal data, (ii) the reasoning capabilities needed for detecting valid relationships between situations, and (iii) the operators necessary for a supporting situation-based reasoning. Our investigations are based on a new state concept, which comprises interval timestamped data derived from observed events and processed using the sequenced semantics. States provide a high level description of past, current and future situations defined for (real world) objects to be monitored. For efficiently processing states in a database system, new set operators are proposed and new auxiliary access paths (i.e., indexes) are discussed which are well-suited for modern NUMA architectures. The proposed database functionalities support the management and analysis of streams of temporal data (e.g., events) resulting from complex monitoring scenarios and thus, provide a valuable basis for situation-aware applications.

A core step in any data analytics pipeline is data cleaning. Data scientists report to spend about 80% of their time on cleaning and other data preparation tasks. Research in data cleaning has provided a variety of approaches to address different data quality problems. Typically, the data scientists has to have some prior knowledge about the dataset and the cleaning routines to correctly select and configure the most fitting approach. We argue that particularly for unknown datasets, it is unrealistic to know the data quality problems upfront and to formulate all necessary quality constraints in one shot. The more realistic approach is to pursue an iterative cleaning process, where the user solves data quality problems step by step. This again poses the challenge of identifying the right sequence of cleaning routines and their configurations.

To overcome this problem, we leverage several techniques from machine learning to build a cleaning workflow orchestrator that learns from cleaning tasks in the past and proposes promising cleaning workflows for a new dataset at hand. In this talk, I discuss our advances in this direction and open technical challenges. In particular, I discuss our solution for estimating the performance of error detection routines to select and aggregate the promising ones and explaining their results.