The Bellman Data Quality Browser

Divesh Srivastava

AT&T Labs – Research
divesh@research.att.com

Keynote Talk Abstract

Data quality is a serious concern in complex industrial-scale databases, which often have thousands of tables and tens of thousands of columns. Commonly encountered problems include missing data (null values), duplicates and default values in columns supposed to treated as keys, data inconsistencies (violation of functional dependencies), and poor quality join paths (lack of referential integrity). Compounding the data quality problems are incomplete and out-of-date metadata about the database and the processes used to populate the database. These problems make the task of analyzing data particularly challenging. To effectively address such problems, we have built the Bellman data quality browser at AT&T. Bellman profiles the database and computes concise statistical summaries of the contents of the database, to identify approximate keys, frequent values of a field (often default values), joinable fields with estimates of join sizes paths, and to understand database dynamics (changes in a database over time). In this talk, I'll describe the technology underlying Bellman and how it is used to help make sense of complex databases.