Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt WS 2013/14

Piggyback Profiling: Metadata for Query Results

Data Profiling

Data profiling is the process of examining some given dataset and collecting statistics about that data. Profiling data is an important and frequent activity of any IT professional and researcher and is necessary for various use-cases. Among the simpler results are **statistics**, such as the number of null values and distinct values in a column, its data type, or the most frequent patterns of its data values. Metadata that are more difficult to compute usually involve multiple columns, such as unique column combinations and functional dependencies. A typical use-case for data profiling examines a large, static dataset, for instance, as retrieved from the Web. Apart from simply supporting a developer in **understanding** the dataset, the determined statistics support schema reverse engineering, integration, and optimization.

Dynamic Data

A yet largely unexplored research area is the profiling of dynamic data, i.e., data that changes over time and data that is only just created as it is profiled. Prime examples for such data are query results, i.e., the virtual relations that are the result of a SQL query on a relational database. Providing **on-the-fly metadata** about query results can support application and database developers, DB-admins, and database developers by relieving them of the typical task of eyeballing a query result for correctness or wrapping the original query with COUNTs and SORTs.

Description

The objective of this project is to extend the open-source DBMS PostgreSQL (implemented in C), to provide metadata for query results with as little processing overhead as possible. The key idea is to **piggyback** on existing operator implementations and to reuse and adapt known metadata of the base tables. For a given query, the DBMS should output the query result, its size, its sortation, some metadata for each column (uniqueness, nulls, min, max, etc.), and some metadata for column combinations (functional dependencies, unique columns combinations).

Possible tasks in this project include:

- Extended **operator model** of open-source DBMS to produce both data and metadata (we shall flexibly define which metadata we want to include)
- Adapted **cost model** for extended operators
- **Algorithmic** or **algebraic extensions** to include specific types of metadata
- **Evaluation** to compare against nesting SQL-query into metadata-queries
- **Extended SQL** language to specify metadata (optional)
- Preparation of **article submission** to major database conference

Contact

Prof. Dr. Felix Naumann: felix.naumann@hpi.uni-potsdam.de