On the Design of Base Tables in the SQL Databases of Some Existing Software

Erki Eessaar

Department of Software Science,

Tallinn University of Technology,

Estonia

erki.eessaar@taltech.ee

Outline

- Background and goal.
- How we searched the occurrences of database design problems.
- The results.
  - Statistics.
  - An example.
- Conclusions and future work.

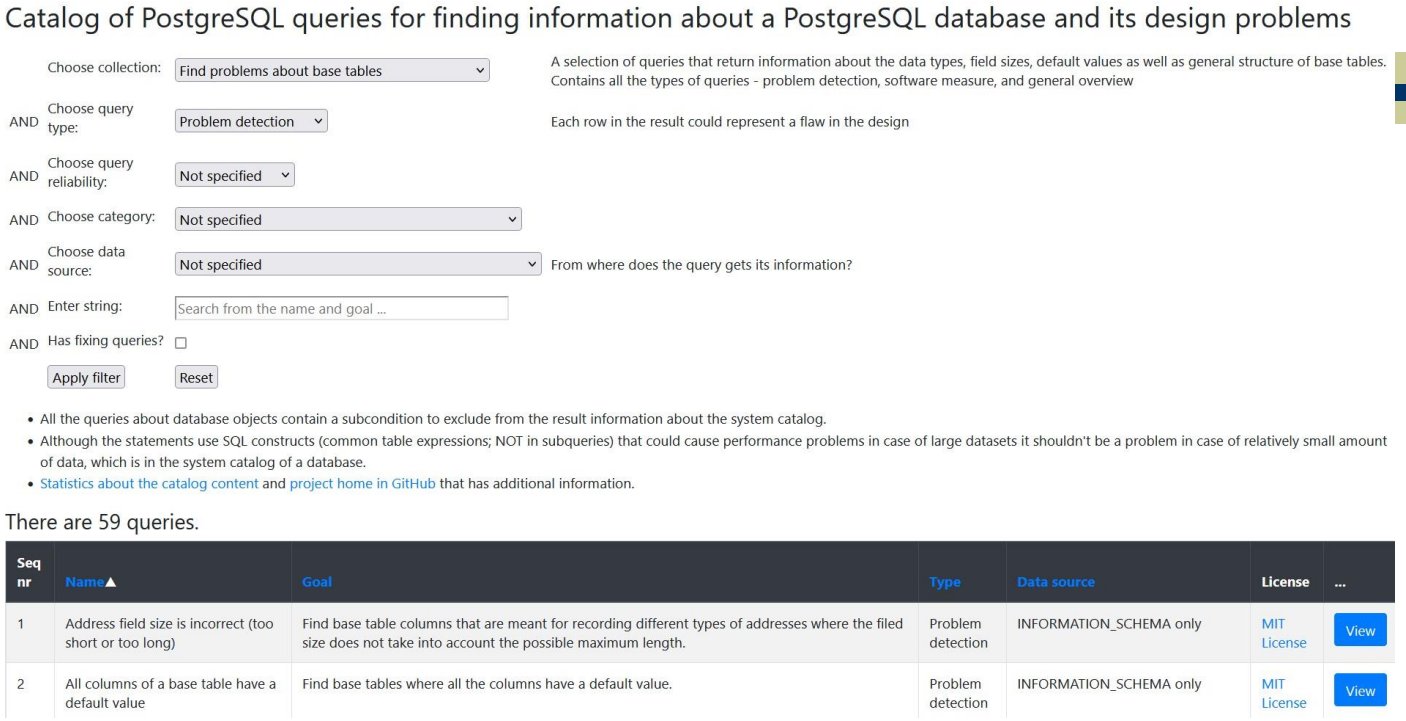
Layered architecture of databases

- Database **conceptual schema** (also known as logical schema) describes the data structures of the entire database for all the potential users of the database.
  - In SQL the data structures are **base tables**.
- The conceptual schema elements are the basis for defining external schemas and hide the details of internal storage (internal schema).

The goal

- Investigate how well the base tables are designed in existing SQL databases.
  - Existing case studies are old or report a small number of problems.
  - *Exclude* the **integrity constraints** – primary key, unique, not null, check.
    - We have analyzed the problems with these before.
  - *Include* the usage of **data types** and **field sizes**, which are also **integrity constraints**.

The catalog



Catalog of PostgreSQL queries for finding information about a PostgreSQL database and its design problems

Choose collection: Find problems about base tables

A selection of queries that return information about the data types, field sizes, default values as well as general structure of base tables. Contains all the types of queries - problem detection, software measure, and general overview

AND Choose query type: Problem detection

Each row in the result could represent a flaw in the design

AND Choose query reliability: Not specified

AND Choose category: Not specified

AND Choose data source: Not specified

From where does the query gets its information?

AND Enter string: Search from the name and goal ...

AND Has fixing queries? ☐

Apply filter Reset

- All the queries about database objects contain a subcondition to exclude from the result information about the system catalog.
- Although the statements use SQL constructs (common table expressions; NOT in subqueries) that could cause performance problems in case of large datasets it shouldn't be a problem in case of relatively small amount of data, which is in the system catalog of a database.
- Statistics about the catalog content and project home in GitHub that has additional information.

There are 59 queries.

Seq nr	Name▲	Goal	Type	Data source	License	...
1	Address field size is incorrect (too short or too long)	Find base table columns that are meant for recording different types of addresses where the filed size does not take into account the possible maximum length.	Problem detection	INFORMATION_SCHEMA only	MIT License	View
2	All columns of a base table have a default value	Find base tables where all the columns have a default value.	Problem detection	INFORMATION_SCHEMA only	MIT License	View

https://github.com/erki77/database-design-queries

We have developed a large set of PostgreSQL system catalog-based queries for searching database design problems of PostgreSQL databases.

The catalog (2)

- Many of the queries directly point to problem occurrences.
  - Mistakes.
  - Design smells.
    - Will cause later maintenance problems.
- Each such query documents a design problem.
  - The absolute majority of these could appear in the databases of any SQL DBMS.

- A long development history, still actively used
- Use a PostgreSQL database

The analysis – databases

- FusionForge
  - An open source development management and team collaboration software.
  - Development started in **2001**.
  - **206** base tables (tables from now on) and **1097** columns.
- LedgerSMB
  - An open source enterprise resource planning software.
  - Development started in **2006**.
  - **162** tables and **978** columns.

The analysis – databases (2)

- OTRS Community Edition
  - An open source ticketing software, which can be used to track and manage issues that need resolving.
  - Development started in **2001**.
  - **116** tables and **962** columns.
- Stansoft
  - A Linux financial accounting software.
  - **174** tables and **1931** columns.

Resulting catalog of problems

- In total, we identified **31** problems in the analyzed databases.
- Many have *more than one* **sign**, i.e., subproblems.
  - The collection of the used queries for this research contains **59** *problem detection queries*.
- **7** problems were present in all the databases.
  - **60%** (**3**) of all the identified data type problems.

Resulting catalog of problems (2)

- We searched the occurrences of a larger set of problems (**151** problem detection queries).
- We present only the problems that had at least occurrence in at least one of the databases.
- We did not find any literature reference to **55%** (**17**) of these problems.

A classification of the problems (problem area)

- Data types of columns
  - **5** problems
    - **1** not in the literature
- Field sizes
  - **5** problems
    - **3** not in the literature
- Default values
  - **5** problems
    - **None** in the literature
- Structure of base tables
  - **13** problems
    - **5** not in the literature
- Sequence generators
  - **3** problems
    - **None** in the literature

A classification of the problems (problem reason)

- Something is missing
  - **7** problems. For instance, maximum field sizes, default values, and classifier tables.
- Inconsistencies
  - **6** problems. For instance, in using data types, field sizes, and default values.
- Imprecision
  - **5** problems. For instance, in determining field sizes.

A classification of the problems (problem reason) (2)

- Unneeded elements
  - **3** problems. For instance, columns for durations.
- Incorrectness
  - Structure of base tables
    - **6** problems. For instance, many SQL database design antipatterns form the seminal book of B. Karwin.
  - Column properties
    - **4** problems. For instance, using a floating-point type (rounding errors) or an incorrect data type.

Karwin, B.: *SQL Antipatterns*. Avoiding the Pitfalls of Database Programming. The Pragmatic Bookshelf (2010)

Data types
Inconsistencies

An example

- In different tables the columns with the same name had different types.
  - The number of such column names: LedgerSMB **20**, FusionForge **13**, OTRS **8**, and Stansoft **6**.
  - In FusionForge column name *type* had **4** different data types – TEXT, CHAR, VARCHAR, and INTEGER.
    - FusionForge had the biggest number of different problems present (**28**).

Conclusions

- All the databases had a lot of different design problems regarding base tables.
- We presented a lot of design problems that have not been published before.
- Most of the problems do not prevent the usage of the database.
  - Will cause difficulties in understanding and maintaining the databases, i.e., **technical debt**.

Future work

- Investigating other databases in terms of the same problems.
  - Perhaps development practices of **commercial systems** lead to *different outcomes*.
- Investigating external schemas, internal schema, and naming of database objects in the databases of existing programs

Thank you for your attention!

Questions?

Erki Eessaar – erki.eessaar@taltech.ee

Reference to the catalog:
https://github.com/erki77/database-design-queries

- Collection “Find problems about base tables”.