CeADAR
Centre for Applied Data Analytics Research

Data Management for Analytics

Unfamiliar Data Modeller

MARKET NEED

ETL (Extract Transform Load) software systems are responsible for the extraction of data from various sources, the cleansing, validation, and reformatting of data, and the insertion of data into a data warehouse.

In some cases the source data is unfamiliar to the developer or analyst and has no apparent structure. ETL tools typically do not try to infer relationships between source entities, with the result that the user is forced to configure filters, joins, aggregations manually to define the transformation from source to destination. Doing so can be time-consuming and error-prone.

TECHNOLOGY SOLUTION

In this project we address the need to automate at least part of that manual process, as applied to source data obtained from relational databases.

The task of delivering the required functionality is divided into two parts. The first is the discovery by inference of any implicit structure within an unfamiliar data set. The second is the presentation of the results of that inference to the user, allowing them to update the metadata associated with the data set such that inferred relations are explicitly declared.

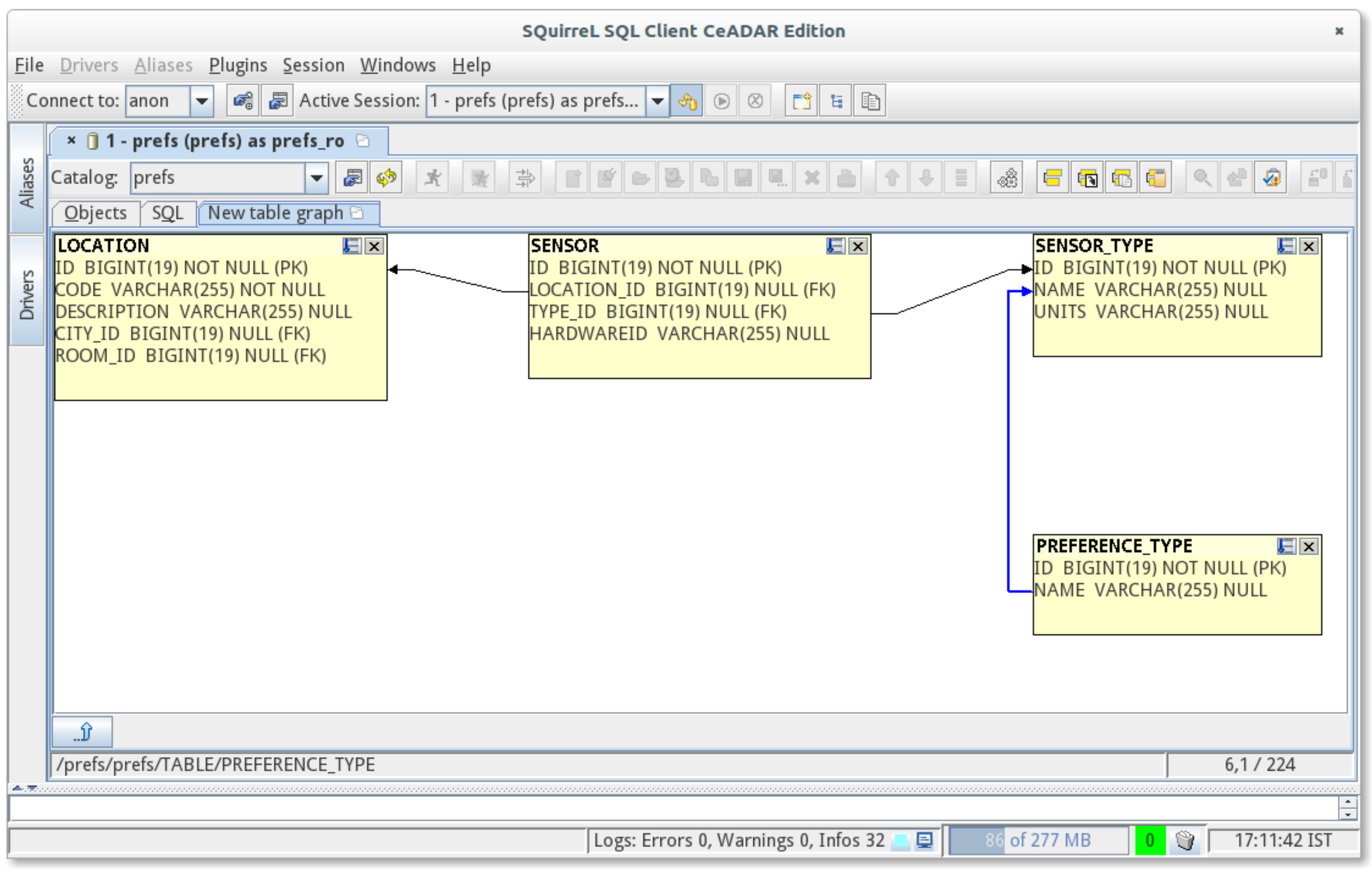
1. $\forall k_i, k_j \in K, k_i \neq k_j$
2. $\forall k \in K, k \neq \emptyset$
3. $Type(K) = Type(R)$
4. $R \subset K$
5. $Min(R) \geq Min(K)$
6. $Max(R) \leq Max(K)$
7. $|R| \leq |K|$

Database relationship tests. We let K *be the set of keys and* R *the set of reference values, having discarded any null values in the latter. If all of the conditions are satisfied then we infer a relationship between the key and reference relations. Otherwise no relationship is inferred.*

USER INTERFACE

The user interface is based on the SQuirreL SQL Client, a graphical SQL client which allows the user to view the structure of a database, browse the data in tables, issue SQL commands, SQuirrel is written in Java, and supports any JDBC-compliant database.

As well as offering a rich set of features, SQuirreL allows new features to be added by means of its Plugin API. Furthermore, SQuirreL provides a standard mechanism for plugins to communicate with one another. We take advantage of both of those capabilities by (a) writing a new plugin which invokes the inference algorithms and (b) using the existing 'Graph' plugin and core SQuirreL libraries to illustrate the results.



Connect to: anon
Active Session: 1 - prefs (prefs) as prefs...
1 - prefs (prefs) as prefs_ro
Catalog: prefs

LOCATION
ID BIGINT(19) NOT NULL (PK)
CODE VARCHAR(255) NOT NULL
DESCRIPTION VARCHAR(255) NULL
CITY_ID BIGINT(19) NULL (FK)
ROOM_ID BIGINT(19) NULL (FK)

SENSOR
ID BIGINT(19) NOT NULL (PK)
LOCATION_ID BIGINT(19) NULL (FK)
TYPE_ID BIGINT(19) NULL (FK)
HARDWAREID VARCHAR(255) NULL

SENSOR_TYPE
ID BIGINT(19) NOT NULL (PK)
NAME VARCHAR(255) NULL
UNITS VARCHAR(255) NULL

PREFERENCE_TYPE
ID BIGINT(19) NOT NULL (PK)
NAME VARCHAR(255) NULL

/prefs/prefs/TABLE/PREFERENCE_TYPE 6,1 / 224

Logs: Errors 0, Warnings 0, Infos 32 86 of 277 MB 0 17:11:42 IST

The Unfamiliar Data Modeller, showing an inferred relation in blue.

RESEARCH TEAM

The research team at UCC consists of Helmut Simonis, Liam O'Toole, and Dhani Merrick

TECHNOLOGY CENTRE
ENTERPRISE IRELAND
IDA IRELAND SUPPORTED

www.ceadar.ie +353 (0)1 716 5716