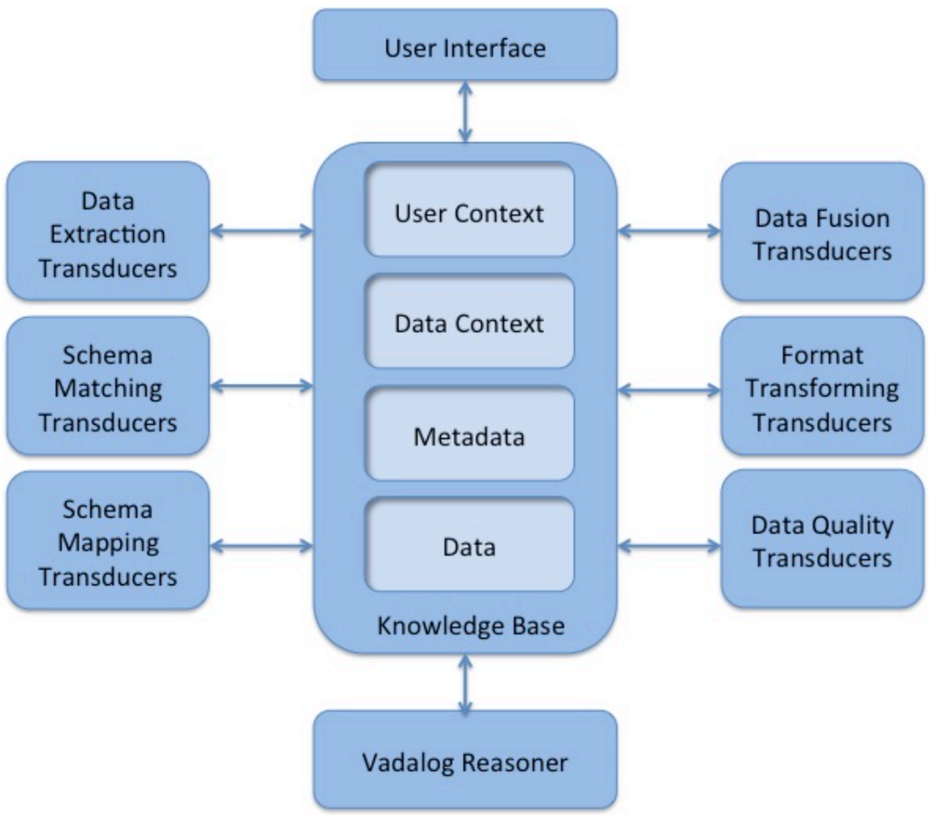
The VADA Architecture for Cost-Effective Data Wrangling



End-To-End Data Wrangling

The VADA architecture: (i) combines wrangling components that between them cover the complete data wrangling lifecycle; (ii) builds on automation wherever possible, using whatever information is available; (iii) refines the results of automated processes in the light of user feedback; and (iv) takes into account the user's priorities.

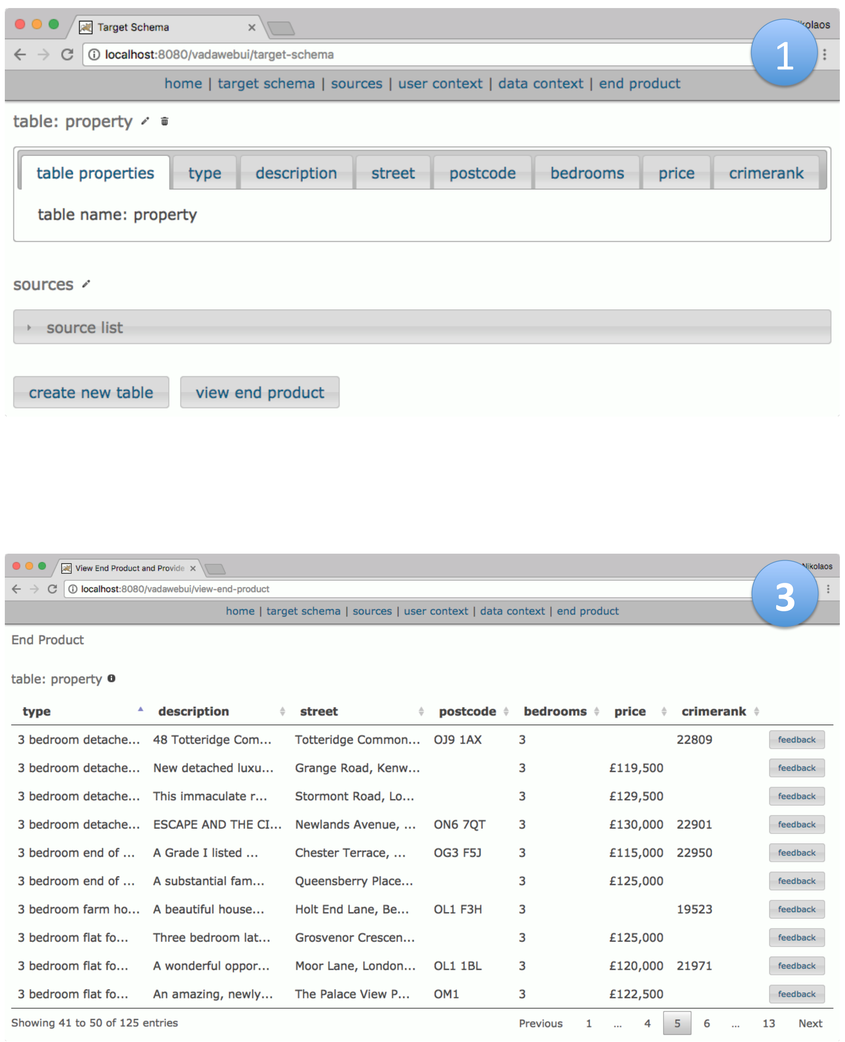
User Interface: The data scientist provides information about the data they need, their priorities and feedback.

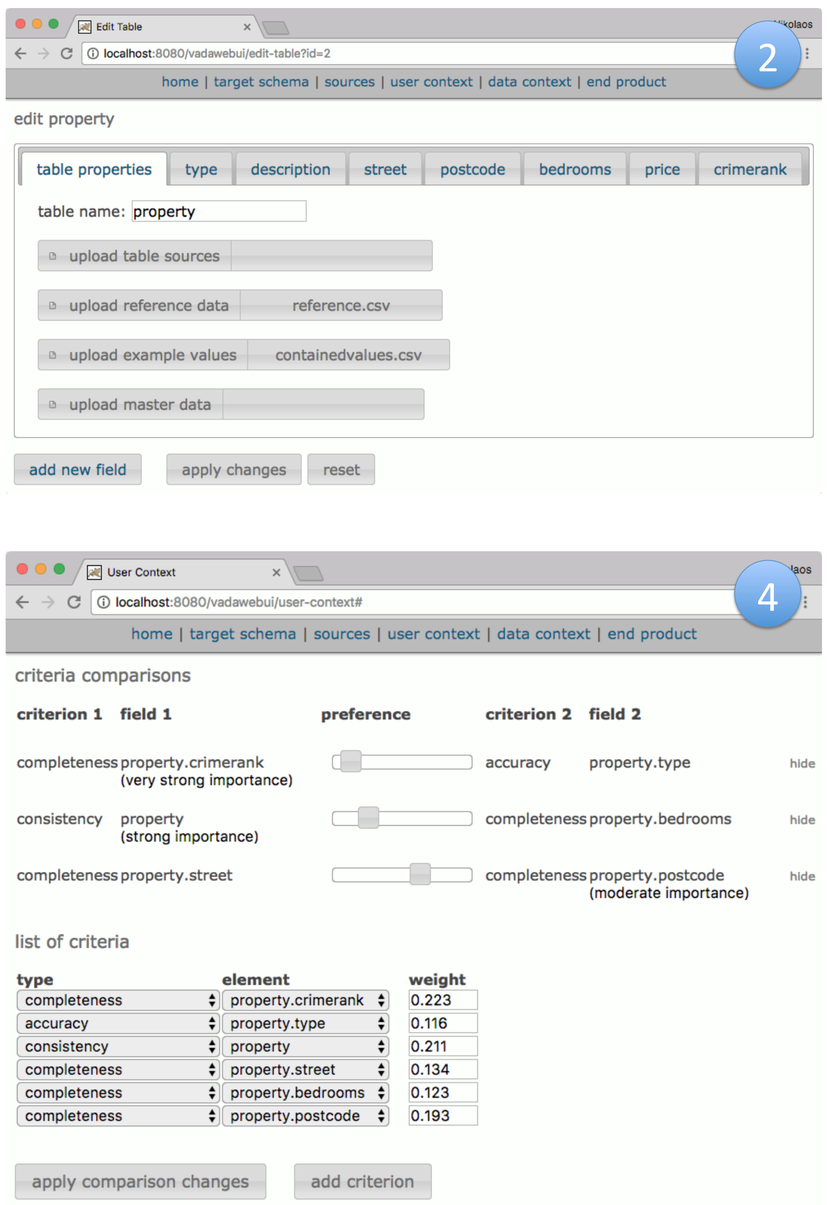
Transducers: Components with input and output dependencies defined as Datalog$^{\pm}$ rules over the knowledge base.

User Context: Information about the requirements of the user.

Data Context: Information about the application domain.

Vadalog Reasoner: Supports reasoning over the knowledge base using Vadalog, a member of the Datalog$^{\pm}$ family of languages.





Demonstration

1. Automatic Bootstrapping: The user identifies a collection of sources and defines a target schema. The system automatically orchestrates a collection of transducers that together generate an initial result data set.

2. Data context: The user associates the target schema with related extents (e.g. master data). Such data allows various of the steps from bootstrapping to be revisited, including matching and mapping validation. Furthermore, it is now also possible to learn quality rules, and thereby to carry out repairs to the mapping results. The result data should now be of better quality.

3. Feedback: The user provides feedback to indicate that some of the results are correct or incorrect. Depending on the feedback provided, this will enable some of the previous steps in the wrangling process to be revisited, giving rise to a revised result.

4. User context: The result is now hopefully of reasonable quality, but the data included in the result may not be especially well-suited to the task at hand. The user specifies the user context by indicating the relative (pairwise) importance of different features in the result. The pairwise comparisons are used to derive weights that inform the selection of mappings based on multi-dimensional optimization.

A Real Estate Scenario

The demonstration acts on real estate data, bringing together:

- Data about properties for sale from web data extraction over deep web sources.
- Open government data that provides information about the areas in which the properties are located.

The University of Manchester
UNIVERSITY OF OXFORD
THE UNIVERSITY of EDINBURGH

EPSRC VADA: Value Added Data Systems – Principles and Architecture www.vada.org.uk