Percipient

intelligent data engineering

UniRefine

Interactive Wrangling of Unified Data (Beta)

A Percipient Technology White Paper

Author: Ravi Shankar Nair

Chief Technology Officer

Updated Nov 2017

Overview

Data wrangling is the process of cleaning messy and noisy datasets, typically for the purposes of data analytics and reporting.

While there are many data wrangling tools in the market, UniRefine is unique in enabling users to interactively investigates ample chunks of pre-aggregated data sourced from multiple systems and formats.

This data is searched to understand patterns and irregularities, or to identify data that matches a particular criteria. The user is then able to apply relevant transformations to the full, unified dataset, as part of the process to prepare data for easy discovery and mining. Other tools will tend to allow investigations and transformations only on single-sourced data.

What can UniRefine Do?

UniRefine facilitates quick and intuitive assessment of manageable amounts of data aggregated from heterogeneous sources, systems or locations. This can include data external to an enterprise's environment.

It is an add-on tool to Percipient's UniConnect data integration platform but is sufficiently lightweight to be used independently as a desktop application, without the need to upload sensitive data to the Web.

UniRefine also does not requires users to have deep technical know-how. Instead, it is designed for data professionals interested in exploring and aligning datasets for downstream application, such as analytics, reporting and visualisation.

> Unirefine is designed for data professionals interested in exploring and aligning datasets for downstream application, such as analytics, reporting and visualisation.

To do this, UniRefine helps users:

- ✓ import pre-aggregateddata
- ✓ facet and filter data
- ✓ determine similar characteristics across data column/ rows
- ✓ cluster together common data cells/ columns/ rows
- ✓ detect inconsistencies across data cells/ columns/ rows
- ✓ change data cells or columns/ rows in bulk using a single command
- ✓ experiment with transforming data from one format to another
- ✓ align names of data fields to name registries (databases)

UniRefine is able to leverage on a familiar spreadsheet interface to ease the process of data investigation. It does not require a schema and is based on a tabular format. However, the tool includes some important features not offered by spreadsheets

UniRefine	
Units of interaction	Columns and rows are the primary units of interaction
Input	Designed for inputting and assessing complex and disparate datasets
Transparency	Each step of the data transformation is shown
Exploration	Inbuilt mechanisms for exploring and understanding the data
Identification	Use of facets and filters to isolate patterns in the data
Integration	Integrated with the UniConnect data integration platform

Scalability & Traceability

By virtue of its integration with the UniConnect data integration platform, UniRefine enables users to import and view enterprise data sourced from disparate systems.

UniRefine is designed to handle large but manageable amounts of data. For example, the tool can comfortably wrangle 100, 000 rows by 10 columns. Users can experiment with cleaning processes on this large sample size, by retaining the ability to test, undo and redo edits as required.

Another key feature of the UniRefine tool is its ability to maintain a history of data cleaning actions taken, and to export this history if required.

Easy To Use

> UniRefine is able to maintain a history of data cleaning actions taken, and to export this history if required.

There are five typical steps to using UniRefine:

Step 1: Launch

You can use your web browser to launch UniRefine

Step2: Import

You can choose to use UniRefine for data that is already stored on your desktop, available via a private web address, or is publicly available on Google. Alternatively, for data residing in your enterprise's databases, the data can be fetched from UniConnect.

The UniConnect option simply requires you to write a query, then paste the results into the UniRefine clipboard in any of the supported formats: TSV, CSV, Excel, JSON, XML, RDF as XML.

Step 3: Facet & Filter

The imported data can then be explored using facets. These text or numeric facets helps you to start to understand selective records within your dataset. By applying a filter, the records displayed can be further refined. Filters can be removed or reset as required

Step 4: Custom Facet

A custom facet such as a text length facet, or a numeric log facet can then be applied, allowing you to start to parse data and identify inconsistencies.

Step 5: Cell / Row / Column Editing

At this point, there are several ways to transform inconsistent data found in individual or multiple cells, rows or columns. You can, for example, choose to transform cells using expressions, clustering, or correct field overloading.

UniRefine further enables data to be split into several columns or new columns added, while duplicate rows can be deleted. It is also possible to conduct structural editing, ie either transposing either a fixed or variable number of columns into rows, or rows into columns.

Use Cases

In this way, UniRefine allows users to undertake a range of data wrangling use cases.

For example, the tool lets users assess how to implement a consistent data vocabulary. By assessing how closely a field conforms to a particular schema, users can interactively align all identical cells or rows.

UniRefine also lets users identify matched data and then test how these can be linked.

UniRefine also lets users identify matched data and then test how these can be linked. By searching for a match across items in a field, users can seek to cluster together related data. Similarly, related data can be combined into single row records. This is a way to convert complex records into more digestible data formats.

Once users understand how the sample data can be sorted, clustered, deleted or changed, these transformation can be executed by leveraging UniConnect's range of in-built SQL functions. By transforming and cleaning the multi-sourced data joined and made accessible via the UniConnect interface, users are able to achieve more reliable analytics and business intelligence.