Lenses: An On-Demand Approach to ETL

Ying Yang*, Niccolò Meneghetti*, Ronny Fehling*, Zhen Hua Liu*, Oliver Kennedy*
+ SUNY Buffalo, * Oracle
{yyang25, niccolom, okennedy}@buffalo.edu
{ronny.fehling, zhen.liu}@oracle.com

The ODIn Lab @ UB
University at Buffalo
The State University of New York™

Motivation
Efficient analytics depends on accurate, reliable, high-quality information. But, raw data are messy.

1. Upfront cleaning: clean all messy data before analysis. Drawbacks: Unnecessary processing of unused data.
2. Inline cleaning: clean all messy data when analysis. Drawbacks: (1) Unnecessary processing of unused data. (2) Duplication of work.
3. On-demand cleaning: delay the cleaning process until needed and clean incrementally. Advantages: Time and cost efficient compared to 1 and 2. We need a general on-demand cleaning framework.

Example
Alicia is an analyst from HappyBuy. She wants to explore the ratings of HappyBuy products.

Current On-Demand Frameworks
Current systems focus on different data quality problems and have different representations and different forms of feedback.

We need a unified system for on-demand ETL!

Contributions
We propose Mimir to provide:
- Lens: a structure to represent different kinds of messy data in a uniform way.
- Analysis: presenting (uncertain) query results to user.
- Feedback: improving the data quality in a cost efficient way.

Lenses
Lenses make best use of source data and make a best-effort guess using the learnt model.

Cells in a generalized C-Table can have arbitrary expressions.

Domain repair lens
CREATE LENS SaneProduct AS
SELECT p.pid, p.category, r.rating, r.review_ct
FROM Product p, Rating r
WHERE p.category IN ('phone', 'TV')
AND r.rating > 4

Lens and are closed on relational algebra.

Anlysis
Query: SELECT p.pid, p.category, r.rating, r.review_ct
FROM SaneProduct AS
WHERE p.category IN ('phone', 'TV')
AND r.rating > 4

The best effort result hides uncertainty from the user, leaving the user the summary of which results are uncertain.

Feedback
We use entropy to measure how uncertain the best effort result is.

Why uncertain?

Entropy H= 0.5

E[H] = 0.8

How bad is the result?

E[H] = 0.485

We use entropy to rank the uncertainties.