NADEEF: A Commodity Data Cleaning System

Data analytics, QCRI

Michele Dallachiesa  
University of Trento

Amr Ebaid  
Purdue University

Ahmed Eldawy  
University of Minnesota

Ahmed Elmagarmid  
Ihab F. Ilyas  
Mourad Ouzzani  
Nan Tang
Bob should be standardized to Robert
Country code determines a country.
## A Motivating Scenario

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>phn</th>
<th>when</th>
<th>where</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>1pm 6/05/2012</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>11am 2/12/2011</td>
<td>Netherlands</td>
</tr>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>Netherlands</td>
<td>66700541</td>
<td>6am 6/05/2012</td>
<td>US</td>
</tr>
<tr>
<td>Paul</td>
<td>Market</td>
<td>Amsterdam</td>
<td>31</td>
<td>UK</td>
<td>55384922</td>
<td>9am 6/02/2012</td>
<td>Netherlands</td>
</tr>
</tbody>
</table>

### Bank Transactions

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>tel</th>
<th>gd</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>M</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>M</td>
</tr>
</tbody>
</table>
## A Motivating Scenario

### tran

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>phn</th>
<th>when</th>
<th>where</th>
</tr>
</thead>
<tbody>
<tr>
<td>David Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>Netherlands</td>
<td>66700543</td>
<td>1pm 6/05/2012</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Paul Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>Netherlands</td>
<td>44944631</td>
<td>11am 2/12/2011</td>
<td>Netherlands</td>
</tr>
<tr>
<td>David Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>Netherlands</td>
<td>66700541</td>
<td>6am 6/05/2012</td>
<td>US</td>
<td></td>
</tr>
<tr>
<td>Paul Market</td>
<td>Amsterdam</td>
<td>31</td>
<td>UK</td>
<td>55384922</td>
<td>9am 6/02/2012</td>
<td>Netherlands</td>
<td></td>
</tr>
</tbody>
</table>

### bank

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>tel</th>
<th>gd</th>
</tr>
</thead>
<tbody>
<tr>
<td>David Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>Paul Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>M</td>
<td></td>
</tr>
</tbody>
</table>

If a customer’s CC is 31, but his/her country is neither Netherlands nor Holland, update the country to Netherlands;

**ETL rules (lookup table)**

Extended CFDs
### A Motivating Scenario

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>phone</th>
<th>when</th>
<th>where</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>1pm 6/05/2012</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>11am 2/12/2011</td>
<td>Netherlands</td>
</tr>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>Netherlands</td>
<td>66700541</td>
<td>6am 6/05/2012</td>
<td>US</td>
</tr>
<tr>
<td>Paul</td>
<td>Market</td>
<td>Amsterdam</td>
<td>31</td>
<td>UK</td>
<td>55384922</td>
<td>9am 6/02/2012</td>
<td>Netherlands</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>tel</th>
<th>gd</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>M</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>M</td>
</tr>
</tbody>
</table>

**Editing rules (w.r.t. master data)**

If the same person from different tables has different phones, the phone number from table bank is more reliable.
## A Motivating Scenario

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>phn</th>
<th>when</th>
<th>where</th>
</tr>
</thead>
<tbody>
<tr>
<td>David Holywell</td>
<td>Oxford</td>
<td>44 UK</td>
<td>44</td>
<td>Netherlands</td>
<td>66700543</td>
<td>1pm 6/05/2012</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Paul Ratcliffe</td>
<td>Oxford</td>
<td>44 UK</td>
<td>44</td>
<td>Netherlands</td>
<td>44944631</td>
<td>11am 2/12/2011</td>
<td>Netherlands</td>
</tr>
<tr>
<td>David Holywell</td>
<td>Oxford</td>
<td>44 Netherlands</td>
<td>44</td>
<td>Netherlands</td>
<td>66700541</td>
<td>6am 6/05/2012</td>
<td>US</td>
</tr>
<tr>
<td>Paul Market</td>
<td>Amsterdam</td>
<td>31 UK</td>
<td>31</td>
<td>Netherlands</td>
<td>55384922</td>
<td>9am 6/02/2012</td>
<td>Netherlands</td>
</tr>
</tbody>
</table>

**bank**

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>tel</th>
<th>gd</th>
</tr>
</thead>
<tbody>
<tr>
<td>David Holywell</td>
<td>Oxford</td>
<td>44 UK</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>M</td>
</tr>
<tr>
<td>Paul Ratcliffe</td>
<td>Oxford</td>
<td>44 UK</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>M</td>
</tr>
</tbody>
</table>

**CFDs (FDs)**

A country code (CC) uniquely determines a country
A Motivating Scenario

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>phn</th>
<th>when</th>
<th>where</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>1pm 6/05/2012</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>11am 2/12/2011</td>
<td>Netherlands</td>
</tr>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>Netherlands</td>
<td>66700541</td>
<td>6am 6/05/2012</td>
<td>US</td>
</tr>
<tr>
<td>Paul</td>
<td>Market</td>
<td>Amsterdam</td>
<td>31</td>
<td>UK</td>
<td>55384922</td>
<td>9am 6/02/2012</td>
<td>Netherlands</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>tel</th>
<th>gd</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>M</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>M</td>
</tr>
</tbody>
</table>

If two purchases of the same person happened in the Netherlands and the US (East Coast) within 1 hour, these two purchases might be a fraud.

Write a special-purpose application
# A Motivating Scenario

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>phn</th>
<th>when</th>
<th>where</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>1pm 6/05/2012</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>11am 2/12/2011</td>
<td>Netherlands</td>
</tr>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>Netherlands</td>
<td>66700541</td>
<td>6am 6/05/2012</td>
<td>US</td>
</tr>
<tr>
<td>Paul</td>
<td>Market</td>
<td>Amsterdam</td>
<td>31</td>
<td>UK</td>
<td>55384922</td>
<td>9am 6/02/2012</td>
<td>Netherlands</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>tel</th>
<th>gd</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>M</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>M</td>
</tr>
</tbody>
</table>

If two purchases of the same person happened in the Netherlands and the US (East Coast) within 1 hour, these two purchases might be a fraud.

Write a special-purpose application

Challenging to capture multiple types of rules
The User Perspective

These are our data quality rules

CFD

MD

Customized rule
The User Perspective

These are our data quality rules

CFD

MD

Customized rule

Data Cleaning System
The User Perspective

These are our data quality rules

CFD

MD

Customized rule

Data Cleaning System

Easy to specify and easy to deploy
Challenges

- Heterogeneity
- Interdependency
- Deployment and extensibility
- Metadata management and user interaction
Challenges

• Heterogeneity
• Interdependency
• Deployment and extensibility
• Metadata management and user interaction

• Integrity constraints (CFDs, DCs)
  ETL rules, customized rules
Challenges

- Heterogeneity
- Interdependency
- Deployment and extensibility
- Metadata management and user interaction
- Integrity constraints (CFDs, DCs)
- ETL rules, customized rules
- Interaction of various types of rules
Challenges

• Heterogeneity
• Interdependency
• Deployment and extensibility
• Metadata management and user interaction

• Integrity constraints (CFDs, DCs)
• ETL rules, customized rules
• Interaction of various types of rules
• Download, compile and run
• Extend with new cleaning solutions
Challenges

• Heterogeneity
• Interdependency
• Deployment and extensibility
• Metadata management and user interaction

• Integrity constraints (CFDs, DCs) ETL rules, customized rules
• Interaction of various types of rules
• Download, compile and run Extend with new cleaning solutions
• Dashboard and metadata profiling
NADEEF Architecture

Data Loader

Rule Collector

ETLs, CFDs, MDs, Business rules

Data
NADEEF Architecture

Data Loader

Rule Collector

ETLs, CFDs, MDs, Business rules

Data

Detection and Cleaning Algorithms

Rules

Core

- Rule compiler
- Violation detection
- Data repairing
NADEEF Architecture

metadata management and data custodians

Data Loader

Rule Collector

Data

Metadata

Data Quality Dashboard

Detection and Cleaning Algorithms

extensibility

heterogeneity

interdependency

Core

Rule compiler

Violation detection

Data repairing

Rules

ETLs, CFDs, MDs, Business rules

Users
NADEEF Architecture

metadata management and data custodians

Data Loader

Rule Collector

Data

Metadata

Detection and Cleaning Algorithms

Rules

Data Quality Dashboard

Core

Rule compiler

Violation detection

Data repairing

extensibility

etetogeneity

interdependency

Demonstration at VLDB 2013
NADEEF Architecture

metadata management and data custodians

Data Loader

Rule Collector

ETLs, CFDs, MDs, Business rules

Data

Metadata

Detection and Cleaning Algorithms

Extensibility

heterogeneity

interdependency

Rules

Core

Rule compiler

Violation detection

Data repairing

Data Quality Dashboard

NADEEF

Demo at VLDB 2013

Users
NADEEF Architecture

metadata management and data custodians

Data Loader

Rule Collector

Data

Metadata

ETLs, CFDs, MDs, Business rules

Detection and Cleaning Algorithms

Core

Rule compiler

Violation detection

Data repairing

heterogeneity

interdependency

Users

Demo at VLDB 2013

A commodity data cleaning system

NADEEF

metadata management and data custodians

Data Quality Dashboard

Auditing and lineage

Indices

Probabilistic models

extensibility

Core

Rules

NADEEF

open source initiative

A commodity data cleaning system

6
Programming Interface

Rules

NADEEF
Programming Interface

Rules

NADEEEF

Rule

<table>
<thead>
<tr>
<th>static semantics</th>
<th>dynamic semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>vio(tuple t)</td>
<td>fix(violation V)</td>
</tr>
<tr>
<td>vio(tuple t₁, tuple t₂)</td>
<td></td>
</tr>
</tbody>
</table>
Programming Interface

Rules

NADEEF

Rule

<table>
<thead>
<tr>
<th>static semantics</th>
<th>dynamic semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>vio(tuple t)</td>
<td>fix(violation V)</td>
</tr>
<tr>
<td>vio(tuple t₁, tuple t₂)</td>
<td></td>
</tr>
</tbody>
</table>
Sample Rules

(tran) If a customer’s CC is 31, but his/her country is neither Netherlands nor Holland, update the country to Netherlands.
Sample Rules

**(tran)** If a customer’s CC is 31, but his/her country is neither Netherlands nor Holland, update the country to Netherlands.

Class Rule1 {

set\langle cell\rangle **vio**(tuple t) { /*s in table tran */
    if (t[CC] = 31 ∧ !(t[country] = Netherlands ∨ t[country] = Holland))
        return { t[CC, country]; }
    return ∅;
}

set\langle Expression\rangle **fix** (set\langle cell\rangle V) {
    return { V.t[country] ← Netherlands; }
}
}
**Sample Rules**

**Class Rule1**

```plaintext
set<cell> vio(tuple t) { /* in table tran */
    if (t[CC] = 31 \(\land\) !(t[country] = Netherlands \(\lor\) t[country] = Holland))
        return { t[CC, country]; }
    return \(\emptyset\);
}
```

**static semantics: what is wrong**

```plaintext
set<Expression> fix (set<cell> V) {
    return { V.t[country] \leftarrow Netherlands; }
}
```

---

**Example (tran)** If a customer’s CC is 31, but his/her country is neither Netherlands nor Holland, update the country to Netherlands.
Sample Rules

**Class Rule1**

```c
set⟨cell⟩ vio(tuple t) { /* s in table tran */
    if (t[CC] = 31 ∧ !(t[country] = Netherlands ∨ t[country] = Holland))
        return { t[CC, country]; }
    return ∅;
}

set<Expression⟩ fix (set⟨cell⟩ V) {
    return { V.t[country] ← Netherlands; }
}
```

*(tran)* If a customer’s CC is 31, but his/her country is neither Netherlands nor Holland, update the country to Netherlands.

**static semantics: what is wrong**

**dynamic semantics: possible ways to repair**
Sample Rules

(\textit{tran}) If two purchases of the same person happened in the Netherlands and the US (East Coast) within 1 hour, these two purchases might be a fraud.
Sample Rules

(\text{tran}) If two purchases of the same person happened in the Netherlands and the US (East Coast) within 1 hour, these two purchases might be a fraud.

Class Rule4 {

set\langle\text{cell}\rangle \text{vio}(\text{tuple } t_1, \text{tuple } t_2) \{ /* t_1, t_2 \text{ in table } \text{tran} */
if (t_1[\text{name}] \approx t_2[\text{name}] \land t_1[\text{tel}] = t_2[\text{tel}] \land t_1[\text{where}] = \text{Netherlands} \\ \land t_2[\text{where}] = \text{US} \land | t_1[\text{when}] - t_2[\text{when}] | \leq 1 )
\text{return } \{ t_1[\text{name, tel, where, when}]; t_2[\text{name, tel, where, when}]; \}
\text{return } \emptyset; 
\}
}
Sample Rules

(tran) If two purchases of the same person happened in the Netherlands and the US (East Coast) within 1 hour, these two purchases might be a fraud.

Class Rule4 {

set(cell) vio(tuple t1, tuple t2) { /* t1, t2 in table tran */
  if (t1[name] $\approx$ t2[name] \&\& t1[tel] = t2[tel] \&\& t1[where] = Netherlands
  \&\& t2[where] = US \&\& | t1[when] - t2[when] | <= 1 )
    return { t1[name, tel, where, when]; t2[name, tel, where, when]; }  
  return $\emptyset$;
}

static semantics: what is wrong
NADEEF Extensibility
NADEEF Extensibility
NADEEF Extensibility

“I have a dream.”

Martin Luther King
NADEEF Extensibility
NADEEF Extensibility
NADEEF Extensibility
Inside NADEEF
# Inside NADEEF

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>phn</th>
<th>when</th>
<th>where</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>1pm 6/05/2012</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>11am 2/12/2011</td>
<td>Netherlands</td>
</tr>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>Netherlands</td>
<td>66700541</td>
<td>6am 6/05/2012</td>
<td>US</td>
</tr>
<tr>
<td>Paul</td>
<td>Market</td>
<td>Amsterdam</td>
<td>31</td>
<td>UK</td>
<td>55384922</td>
<td>9am 6/02/2012</td>
<td>Netherlands</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>tel</th>
<th>gd</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>M</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>M</td>
</tr>
</tbody>
</table>
### Inside NADEEF

#### Rule 1
- Rule 2
- Rule 3
- Rule 4

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>phn</th>
<th>when</th>
<th>where</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>1pm 6/05/2012</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>11am 2/12/2011</td>
<td>Netherlands</td>
</tr>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>6am 6/05/2012</td>
<td>US</td>
</tr>
<tr>
<td>Paul</td>
<td>Market</td>
<td>Amsterdam</td>
<td>31</td>
<td>Netherlands</td>
<td>55384922</td>
<td>9am 6/02/2012</td>
<td>Netherlands</td>
</tr>
</tbody>
</table>

**bank**

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>tel</th>
<th>gd</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>M</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>M</td>
</tr>
</tbody>
</table>
## Inside NADEEF

### Rule 1
- Rule 2
- Rule 3
- Rule 4

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>phn</th>
<th>when</th>
<th>where</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>1pm 6/05/2012</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>11am 2/12/2011</td>
<td>Netherlands</td>
</tr>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>6am 6/05/2012</td>
<td>US</td>
</tr>
<tr>
<td>Paul</td>
<td>Market</td>
<td>Amsterdam</td>
<td>31</td>
<td>Netherlands</td>
<td>55384922</td>
<td>9am 6/02/2012</td>
<td>Netherlands</td>
</tr>
</tbody>
</table>

### tran

### bank

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>tel</th>
<th>gd</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>M</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>M</td>
</tr>
</tbody>
</table>
### Inside NADEEF

<table>
<thead>
<tr>
<th>tran</th>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>phn</th>
<th>when</th>
<th>where</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>1pm 6/05/2012</td>
<td>Netherlands</td>
</tr>
<tr>
<td></td>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>11am 2/12/2011</td>
<td>Netherlands</td>
</tr>
<tr>
<td></td>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>6am 6/05/2012</td>
<td>US</td>
</tr>
<tr>
<td></td>
<td>Paul</td>
<td>Market</td>
<td>Amsterdam</td>
<td>31</td>
<td>Netherlands</td>
<td>55384922</td>
<td>9am 6/02/2012</td>
<td>Netherlands</td>
</tr>
</tbody>
</table>

### bank

<table>
<thead>
<tr>
<th>name</th>
<th>street</th>
<th>city</th>
<th>CC</th>
<th>country</th>
<th>tel</th>
<th>gd</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Holywell</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>66700543</td>
<td>M</td>
</tr>
<tr>
<td>Paul</td>
<td>Ratcliffe</td>
<td>Oxford</td>
<td>44</td>
<td>UK</td>
<td>44944631</td>
<td>M</td>
</tr>
</tbody>
</table>
Violation Detection
Violation Detection

- Brute force approach (black-boxes)

| CC | country     | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>44</td>
<td>UK</td>
</tr>
<tr>
<td>r2</td>
<td>44</td>
<td>UK</td>
</tr>
<tr>
<td>r3</td>
<td>44</td>
<td>Netherlands</td>
</tr>
<tr>
<td>r4</td>
<td>31</td>
<td>UK</td>
</tr>
</tbody>
</table>

Violations: (r1, r3), (r2, r3)
Violation Detection

• Brute force approach (black-boxes)
• Optimized approach (white-boxes, e.g., CC->country)

|   | CC | country   | ...
|---|----|-----------|
| r1 | 44 | UK        | ...  
| r2 | 44 | UK        | ...  
| r3 | 44 | Netherlands | ...  
| r4 | 31 | UK        | ...  

Violations: 
(r1, r3), (r2, r3)
Violation Detection

- Brute force approach (black-boxes)
- Optimized approach (white-boxes, e.g., CC->country)

<table>
<thead>
<tr>
<th></th>
<th>CC</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>44</td>
<td>UK</td>
</tr>
<tr>
<td>r2</td>
<td>44</td>
<td>UK</td>
</tr>
<tr>
<td>r3</td>
<td>44</td>
<td>Netherlands</td>
</tr>
<tr>
<td>r4</td>
<td>31</td>
<td>UK</td>
</tr>
</tbody>
</table>

Violations:
(r1, r3), (r2, r3)

**partition**
Violation Detection

- Brute force approach (black-boxes)
- Optimized approach (white-boxes, e.g., CC->country)

<table>
<thead>
<tr>
<th>CC</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>44</td>
</tr>
<tr>
<td>r2</td>
<td>44</td>
</tr>
<tr>
<td>r3</td>
<td>44</td>
</tr>
<tr>
<td>r4</td>
<td>31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CC</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>r12</td>
<td>44</td>
</tr>
<tr>
<td>r3</td>
<td>44</td>
</tr>
<tr>
<td>r4</td>
<td>31</td>
</tr>
</tbody>
</table>

Violations: (r1, r3), (r2, r3)

partition

compression
Violation Detection

- Brute force approach (black-boxes)
- Optimized approach (white-boxes, e.g., CC->country)

<table>
<thead>
<tr>
<th>CC</th>
<th>country</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>44</td>
<td>UK</td>
</tr>
<tr>
<td>r2</td>
<td>44</td>
<td>UK</td>
</tr>
<tr>
<td>r3</td>
<td>44</td>
<td>Netherlands</td>
</tr>
<tr>
<td>r4</td>
<td>31</td>
<td>UK</td>
</tr>
</tbody>
</table>

Violations: (r1, r3), (r2, r3) (r12, r3)

- partition
- compression
Data Repairing
Holistic Data Cleaning

data cleaning

rule specification
Holistic Data Cleaning

<table>
<thead>
<tr>
<th>Violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1: {r4[CC, country]}</td>
</tr>
<tr>
<td>V2: {t1[name, street, city, tel], r3[name, street, city, phn]}</td>
</tr>
<tr>
<td>V3: {r1[CC,country], r3[CC, country]}</td>
</tr>
<tr>
<td>V4: {r2[CC,country], r3[CC, country]}</td>
</tr>
<tr>
<td>V5: {r1[name, tel, where, when], r3[name, tel, where, when]}</td>
</tr>
</tbody>
</table>
### Holistic Data Cleaning

#### Violations

<table>
<thead>
<tr>
<th>V1</th>
<th>{r4[CC, country]}</th>
</tr>
</thead>
<tbody>
<tr>
<td>V2</td>
<td>{t1[name, street, city, tel], r3[name, street, city, phn]}</td>
</tr>
<tr>
<td>V3</td>
<td>{r1[CC, country], r3[CC, country]}</td>
</tr>
<tr>
<td>V4</td>
<td>{r2[CC, country], r3[CC, country]}</td>
</tr>
<tr>
<td>V5</td>
<td>{r1[name, tel, where, when], r3[name, tel, where, when]}</td>
</tr>
</tbody>
</table>

#### Candidate fixes

<table>
<thead>
<tr>
<th>F1</th>
<th>r4[country] ← Netherlands</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2</td>
<td>r3[phn] ← t1[tel]</td>
</tr>
<tr>
<td>F3</td>
<td>r1[country] ← r3[country]</td>
</tr>
<tr>
<td>F4</td>
<td>r3[country] ← r1[country]</td>
</tr>
<tr>
<td>F5</td>
<td>r2[country] ← r3[country]</td>
</tr>
<tr>
<td>F6</td>
<td>r3[country] ← r2[country]</td>
</tr>
</tbody>
</table>
A Variable-Weighted Max-SAT

Violations

Candidate fixes
A Variable-Weighted Max-SAT

Violations → CNF → Candidate fixes
A Variable-Weighted Max-SAT

Violations

Candidate fixes

CNF

Variable-weighted
MAX-SAT solver
A Variable-Weighted Max-SAT

Violations → CNF → Variable-weighted MAX-SAT solver → Repairs

Candidate fixes
A Variable-Weighted Max-SAT

Violations

Candidate fixes

CNF

Variable-weighted MAX-SAT solver

Repairs

variable

$x_{r4}^{Netherlands}$ [country]
A Variable-Weighted Max-SAT

Violations → CNF → Variable-weighted MAX-SAT solver → Repairs

Candidate fixes

variable inclusive assignment

\( x_{r4}[\text{country}] \)

\( (x_{UK}^{\text{country}} \lor x_{Netherlands}^{\text{country}}) \)
A Variable-Weighted Max-SAT

Violations → CNF → Variable-weighted MAX-SAT solver → Repairs

Candidate fixes

variable

inclusive assignment

exclusive assignment

\[ x_{\text{Netherlands}}^{\text{country}} \]

\[ (x_{r4}^{\text{UK}}[\text{country}] \lor x_{r4}^{\text{Netherlands}}[\text{country}]) \]

\[ (\neg x_{r4}^{\text{UK}}[\text{country}] \lor \neg x_{r4}^{\text{Netherlands}}[\text{country}]) \]
A Variable-Weighted Max-SAT

Violations \rightarrow \text{CNF} \rightarrow \text{Variable-weighted MAX-SAT solver} \rightarrow \text{Repairs}

Candidate fixes

variable

inclusive assignment

exclusive assignment

avoid violations

\begin{align*}
\forall \text{r}_4[^\text{country}] \\
(x_{\text{UK}}[^\text{country}] \lor x_{\text{Netherlands}}[^\text{country}]) \\
(x_{\text{UK}}[^\text{country}] \lor \neg x_{\text{r}_4[^\text{country}]}) \\
(\neg x_{\text{31}}[^\text{CC}] \lor \neg x_{\text{UK}}[^\text{country}])
\end{align*}
Experimental Study
Effectiveness

(a) Hospital dataset
(100K, 9 attributes, 10 rules)

(b) Bus dataset
(160K, 16 attributes, 11 rules)
Conclusion

- A generalized programming interface (heterogeneity)
- Holistic data cleaning (interdependency)
- An extensible system (extensibility)