Design Motivation and Impact on the Database Community

Jie Song
Computer Science Engineering
University of Michigan
From Language Conversion To Data Transformation Conversion

- Inter-conversion between each of the languages are expensive, $O(n^2)$
- Bridge by an intermediate language, $O(n)$
From Language Conversion To Data Transformation Conversion

• Traditional general programming language conversion
  • A simple syntax replacing system
  • More recent works begin to consider different levels of conversion
    • E.g., convert code of a procedural language into a purely object oriented language

• Statistical programming languages
  • Simpler programming languages that are mostly procedural
  • R, Python are more general purpose while SPSS, STATA and SAS are limited
  • Focus on data transformation and analysis
From Language Conversion To
Data Transformation Conversion

• Data transformation as a directed acyclic graph
  • Data as the node and transformation as the edge

• Semantic level conversion
  • Regardless of the underlying syntax used by hiding implementation details
  • From higher level, e.g. data set level, to lower level, e.g., column/row level, metadata level

• Graphical Visualization for better understanding
  • Intermediate result when processed with the raw data
  • Data lineage or provenance
  • Code association with transformations

R: D1$V3=D1$V2+1
  for (i in 1:nrow(D1))
    D1$V3[i] = D1$V2[i]+1

Stata: generate V3=V2+1

V4=Count(V1)
groupby V1

D1
<table>
<thead>
<tr>
<th>V1</th>
<th>V2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

D2
<table>
<thead>
<tr>
<th>V1</th>
<th>V2</th>
<th>V3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

D3
<table>
<thead>
<tr>
<th>V1</th>
<th>V2</th>
<th>V3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

D4
<table>
<thead>
<tr>
<th>Index</th>
<th>V2</th>
<th>V3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

D5
<table>
<thead>
<tr>
<th>V1</th>
<th>V4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
The Future of $C^2$ Metadata

- A systematic guideline for building converter that maps from a new statistical language to SDTL
  - Phrase-based statistical translation
- Promote research in
  - Data services, e.g. data sharing and data publication lineage construction
  - Transformation code transfer and reuse
  - Data preparation such as
    - data exploration
    - data quality evaluation
    - data integration
References

