Bootstrapping Pay-As-You-Go Data Integration Systems
Anish Das Sarma, Xin Dong, Alon Halevy

Christian Grant
cgrant@cise.ufl.edu

University of Florida

October 10, 2011
very technical, don’t hate me
Outline

1. Background
2. Goal
3. Motivation
4. Example
5. Overview
6. Definitions
7. Generation
8. Consolidation
9. Experiments
10. Conclusion
11. Thank you
Data integration systems offer a single-point interface over data sources and possibly heterogeneous data.

- Represented by the triple \( \{ G, S, M \} \)
  - \( G \) is the Global schema
  - \( S \) is the set of source schemas
  - \( M \) is the mapping between \( G \) and \( S \)

Two main approaches:

1. **Global As View** – \( G \) is a collection of mediated views
2. **Local As View** – All local views \( S \) are changed to look like \( G \)
Global As View

- **Simple interface** that must be designed in advance
- **Mapping functions** must be created for each source to global
- Queries are reformulated over query to reach data sources
- Results are retrieved and combined from data sources
- Does not scale
Goal

- Give best effort answers over data sources while allowing the administrator improve system in a pay-as-you-go fashion.
Motivation

- Many approaches automatically performed the mappings
- Several plausible mappings may exist 😞
- Idea: Let's make the mappings probabilistic and choose the mapping with maximum entropy \(^1\)

---

\(^1\)The principle of maximum entropy is a postulate which states that, the probability distribution which best represents the current state of knowledge is the one with largest entropy.
Consider source schema \( S_1 \) and \( S_2 \) both describing people

\[
S_1(\text{name, hPhone, hAddr, oPhone, oAddr})
\]

\[
S_2(\text{name, phone, addr})
\]

Possible mappings:

\[
M_1(\{\text{name}\}, \{\text{phone, hPhone, oPhone}\}, \{\text{addr, hAddress, oAddress}\})
\]

\[
M_2(\{\text{name}\}, \{\text{phone, hPhone}\}, \{\text{oPhone}\}, \{\text{addr, oAddress}\}, \{\text{hAddress}\})
\]

\[
M_3(\{\text{name}\}, \{\text{phone, hPhone}\}, \{\text{oPhone}\}, \{\text{addr, hAddress}\}, \{\text{oA}\})
\]

\[
M_4(\{\text{name}\}, \{\text{phone, oPhone}\}, \{\text{hPhone}\}, \{\text{addr, oA}\}, \{\text{hAddress}\})
\]

\[
M_5(\{\text{name}\}, \{\text{phone}\}, \{\text{hPhone}\}, \{\text{oPhone}\}, \{\text{addr}\}, \{\text{hAddress}\}, \{\text{oA}\})
\]
Example from overview

- Consider source schema $S_1$ and $S_2$ both describing people
  - $S_1(name, hPhone, hAddr, oPhone, oAddr)$
  - $S_2(name, phone, addr)$
  - Possible mappings:
    - $M_1(\{name\}, \{phone, hPhone, oPhone\}, \{addr, hAddress, oAddress\})$
    - $M_2(\{name\}, \{phone, hPhone\}, \{oPhone\}, \{addr, oAddress\}, \{hAddress\})$
    - $M_3(\{name\}, \{phone, hPhone\}, \{oPhone\}, \{addr, hAddress\}, \{oA\})$
    - $M_4(\{name\}, \{phone, oPhone\}, \{hPhone\}, \{addr, oA\}, \{hAddress\})$
    - $M_5(\{name\}, \{phone\}, \{hPhone\}, \{oPhone\}, \{addr\}, \{hAddress\}, \{oA\})$
Example from overview

- Consider source schema $S_1$ and $S_2$ both describing people
  - $S_1(\text{name}, h\text{Phone}, h\text{Addr}, o\text{Phone}, o\text{Addr})$
  - $S_2(\text{name}, \text{phone}, \text{addr})$
  - Possible mappings:
    - $M_1(\{\text{name}\}, \{\text{phone}, h\text{Phone}, o\text{Phone}\}, \{\text{addr}, h\text{Address}, o\text{Address}\})$
    - $M_2(\{\text{name}\}, \{\text{phone}, h\text{Phone}\}, \{\text{oPhone}\}, \{\text{addr}, o\text{Address}\}, \{h\text{Address}\})$
    - $M_3(\{\text{name}\}, \{\text{phone}, h\text{Phone}\}, \{\text{oPhone}\}, \{\text{addr}, h\text{Address}\}, \{o\text{Address}\})$
    - $M_4(\{\text{name}\}, \{\text{phone}, o\text{Phone}\}, \{\text{hPhone}\}, \{\text{addr}, o\text{A}\}, \{h\text{Address}\})$
    - $M_5(\{\text{name}\}, \{\text{phone}\}, \{\text{hPhone}\}, \{\text{oPhone}\}, \{\text{addr}\}, \{h\text{Address}\}, \{o\text{A}\})$

\[ S_1(\text{name}, h\text{Phone}, h\text{Addr}, o\text{Phone}, o\text{Addr}) \]

\[ S_2(\text{name}, \text{phone}, \text{address}) \]

\[ M_2 \]
Example from overview

- Consider source schema $S_1$ and $S_2$ both describing people
  - $S_1(\text{name, hPhone, hAddr, oPhone, oAddr})$
  - $S_2(\text{name, phone, addr})$
  - Possible mappings:
    - $M_1(\{\text{name}\}, \{\text{phone, hPhone, oPhone}\}, \{\text{addr, hAddress, oAddress}\})$
    - $M_2(\{\text{name}\}, \{\text{phone, hPhone}\}, \{\text{oPhone}\}, \{\text{addr, oAddress}\}, \{\text{hAddress}\})$
    - $M_3(\{\text{name}\}, \{\text{phone, hPhone}\}, \{\text{oPhone}\}, \{\text{addr, hAddress}\}, \{\text{oA}\})$
    - $M_4(\{\text{name}\}, \{\text{phone, oPhone}\}, \{\text{hPhone}\}, \{\text{addr, oA}\}, \{\text{hAddress}\})$
    - $M_5(\{\text{name}\}, \{\text{phone}\}, \{\text{hPhone}\}, \{\text{oPhone}\}, \{\text{addr}\}, \{\text{hAddress}\}, \{\text{oA}\})$
Example from overview

Consider source schema $S_1$ and $S_2$ both describing people

- $S_1(\text{name}, \text{hPhone}, \text{hAddr}, \text{oPhone}, \text{oAddr})$
- $S_2(\text{name}, \text{phone}, \text{addr})$

Possible mappings:

$M_1(\{\text{name}\}, \{\text{phone}, \text{hPhone}, \text{oPhone}\}, \{\text{addr}, \text{hAddress}, \text{oAddress}\})$
$M_2(\{\text{name}\}, \{\text{phone}, \text{hPhone}\}, \{\text{oPhone}\}, \{\text{addr}, \text{oAddress}\}, \{\text{hAddress}\})$
$M_3(\{\text{name}\}, \{\text{phone}, \text{hPhone}\}, \{\text{oPhone}\}, \{\text{addr}, \text{hAddress}\}, \{\text{oA}\})$
$M_4(\{\text{name}\}, \{\text{phone}, \text{oPhone}\}, \{\text{hPhone}\}, \{\text{addr}, \text{oA}\}, \{\text{hAddress}\})$
$M_5(\{\text{name}\}, \{\text{phone}\}, \{\text{hPhone}\}, \{\text{oPhone}\}, \{\text{addr}\}, \{\text{hAddress}\}, \{\text{oA}\})$
Example from overview

Consider source schema $S_1$ and $S_2$ both describing people

- $S_1(name, hPhone, hAddr, oPhone, oAddr)$
- $S_2(name, phone, addr)$

Possible mappings:

$M_1(\{name\}, \{phone, hPhone, oPhone\}, \{addr, hAddress, oAddress\})$
$M_2(\{name\}, \{phone, hPhone\}, \{oPhone\}, \{addr, oAddress\}, \{hAddress\})$
$M_3(\{name\}, \{phone, hPhone\}, \{oPhone\}, \{addr, hAddress\}, \{oA\})$
$M_4(\{name\}, \{phone, oPhone\}, \{hPhone\}, \{addr, oA\}, \{hAddress\})$
$M_5(\{name\}, \{phone\}, \{hPhone\}, \{oPhone\}, \{addr\}, \{hAddress\}, \{oA\})$
Example from overview

Consider source schema $S_1$ and $S_2$ both describing people

- $S_1(name, hPhone, hAddr, oPhone, oAddr)$
- $S_2(name, phone, addr)$

Possible mappings:

$M_1(\{name\}, \{phone, hPhone, oPhone\}, \{addr, hAddress, oAddress\})$

$M_2(\{name\}, \{phone, hPhone\}, \{oPhone, addr, oAddress\}, \{hAddress\})$

$M_3(\{name\}, \{phone, hPhone\}, \{oPhone\}, \{addr, hAddress\}, \{oA\})$

$M_4(\{name\}, \{phone, oPhone\}, \{hPhone\}, \{addr, oA\}, \{hAddress\})$

$M_5(\{name\}, \{phone\}, \{hPhone\}, \{oPhone\}, \{addr\}, \{hAddress\}, \{oA\})$

Diagram:

- $S_1(name, hPhone, hAddr, oPhone, oAddr)$
- $S_2(name, phone, address)$
- $M_6$
Consider source schema $S_1$ and $S_2$ both describing people

- $S_1(\text{name}, \text{hPhone}, \text{hAddr}, \text{oPhone}, \text{oAddr})$
- $S_2(\text{name}, \text{phone}, \text{addr})$

Possible mappings:

- $M_1(\{\text{name}\}, \{\text{phone}, \text{hPhone}, \text{oPhone}\}, \{\text{addr}, \text{hAddress}, \text{oAddress}\})$
- $M_2(\{\text{name}\}, \{\text{phone}, \text{hPhone}\}, \{\text{oPhone}\}, \{\text{addr}, \text{oAddress}\}, \{\text{hAddress}\})$
- $M_3(\{\text{name}\}, \{\text{phone}, \text{hPhone}\}, \{\text{oPhone}\}, \{\text{addr}, \text{hAddress}\}, \{\text{oA}\})$
- $M_4(\{\text{name}\}, \{\text{phone}, \text{oPhone}\}, \{\text{hPhone}\}, \{\text{addr}, \text{oA}\}, \{\text{hAddress}\})$
- $M_5(\{\text{name}\}, \{\text{phone}\}, \{\text{hPhone}\}, \{\text{oPhone}\}, \{\text{addr}\}, \{\text{hAddress}\}, \{\text{oA}\})$

- No good mapping exist, even with probabilistic mappings.
- We need both probabilistic mediated schema and semantic mappings to that schema!
Steps

1. Construct a probabilistic mediated schema
2. Find best probabilistic schema mappings
3. Create a single mediated schema to expose to the user
Deterministic Schema Mappings

- We have our source schemas $S = \{S_1, \ldots, S_n\}$
- Attributes of $S_i$ is $\text{attr}(S_i)$
- The set of all source attributes $A = \text{attr}(S_1) \cup \ldots \cup \text{attr}(S_n)$ \(^2\)
- Mediated attributes $M = \{A_1, \ldots, A_n\}$ where $i, j \in [1, m]$, $i \neq j$, $A_i \cap A_j = \emptyset$

\(^2\)Mediated attributes are not given ‘names’
Probablistic Mediated Schema

- Set of schemas $S = \{S_1, \ldots, S_n\}$
- The p-med-schema for $S$ is $M = \{(M_1, Pr(M_1)), \ldots, (M_l, Pr(M_l))\}$ where:
  - $M_i$ is a mediated schema for $S$
  - each $M_i$ is unique
  - $Pr(M_i) \in (0, 1]$, and $\sum_{i=1}^{l} Pr(M_i) = 1$
For a set $S$ and a mediated schema $M$ the probabilistic mapping is

$$pM = \{(m_1, Pr(m_1)), \ldots, (m_l, Pr(m_l))\}$$

- each $m_i$ is a unique schema mapping between $S$ and $M$
- $Pr(m_i) \in (0, 1]$, and $\sum_{i=1}^{l} Pr(m_i) = 1$
Querying Mediated Schema

- Given
  a source schema $S$
  a p-med-schema $M = \{(M1, Pr(M1)), ..., (Ml, Pr(Ml))\}$
  a set of probabilistic mappings $pM = \{pM(M1), ..., pM(Ml)\}$ where each $pM(M_i)$ is a mapping between $S$ and $M_i$.

- Let $D$ be an instance of $S$, $Q$ be a query, and $t$ be a tuple

- The probability $t$ is an answer for $Q$ is $Pr(t|M_i), r \in [1,l]$ w.r.t $M_i$ and $pM(M_i)$.

- Let $p = \sum_{i=1}^{l} Pr(t|M_i) \ast Pr(M_i)$

- We can say $(t, p)$ is a by-table answer w.r.t $M$ and $pM$ if $p > 0$

- the set of table by table answers is $Q_{M,pM}(D)$. 
Consistency

Let $M$ be a mediated schema for sources $S_1, \ldots, S_n$. We say $M$ is consistent with a source schema $S_i, i \in [1, n]$, if there is no pair of attributes in $S_i$ that appear in the same cluster in $M$.

Consistent P-Mapping A p-mapping $pM$ is consistent with a weighted correspondence $C_{i,j}$ between a pair of source and target attributes if the sum of the probabilities of all mappings $m \in pM$ containing correspondence $(i,j)$ equals $p_{i,j}$; that is,

$$p_{i,j} = \sum_{m \in pM, (i,j) \in m} Pr(m).$$

A p-mapping is consistent with a set of weighted correspondences $\mathbf{C}$ if it is consistent with each weighted correspondence $C \in \mathbf{C}$. 
Generate all mediated schema

Algorithm 1: Generate all possible mediated schemas.

0: **Input:** Source schemas $S_1, \ldots, S_n$.
   **Output:** A set of possible mediated schemas.
1: Compute $\mathcal{A} = \{a_1, \ldots, a_m\}$, the set of all source attributes;
2: **for each** $(j \in [1, m])$
   
   Compute frequency $f(a_j) = \frac{|\{i \in [1, n] | a_j \in S_i\}|}{n}$;
3: Set $\mathcal{A} = \{a_j | j \in [1, m], f(a_j) \geq \theta\}$; // $\theta$ is a threshold
4: Construct a weighted graph $G(V, E)$, where (1) $V = \mathcal{A}$, and
   (2) for each $a_j, a_k \in \mathcal{A}$, $s(a_j, a_k) \geq \tau - \epsilon$, there is an edge
   $(a_j, a_k)$ with weight $s(a_j, a_k)$;
5: Mark all edges with weight less than $\tau + \epsilon$ as *uncertain*;
6: **for each** (uncertain edge $e = (a_1, a_2) \in E$)
   
   Remove $e$ from $E$ if (1) $a_1$ and $a_2$ are connected by a
   path with only certain edges, or (2) there exists $a_3 \in V$, such
   that $a_2$ and $a_3$ are connected by a path with only certain edges
   and there is an uncertain edge $(a_1, a_3)$;
7: **for each** (subset of uncertain edges)
   
   Omit the edges in the subset and compute a mediated
   schema where each connected component in the graph corre-
   sponds to an attribute in the schema;
8: **return** distinct mediated schemas.

Algorithm 2: Assign probabilities to possible mediated schemas.

0: **Input:** Possible mediated schemas $M_1, \ldots, M_l$ and source
   schemas $S_1, \ldots, S_n$.
   **Output:** $Pr(M_1), \ldots, Pr(M_l)$.
1: **for each** $(i \in [1, l])$
   
   Count the number of source schemas that are consistent
   with $M_i$, denoted as $c_i$;
2: **for each** $(i \in [1, l])$ Set $Pr(M_i) = \frac{c_i}{\sum_{i=1}^{l} c_i}$.
Figure 3: The p-med-schema for a set of bibliography sources. Each oval in the graph represents an attribute in the mediated schemas. The p-med-schema contains two possible schemas, the first containing attributes in regions A and B₁, and the second containing attributes in regions A and B₂. They have probabilities 0.703 and 0.297 respectively.
We want to choose the mapping that does not favor any of the events over the others

We chose the one that does not introduce new information

1. Enumerate all possible one to one mappings between $S$ and $M$ that have a subset of correspondences $m_1, \ldots, m_l$.

2. Assign probabilities

$$\text{maximize } \sum_{k=1}^{l} -p_k \times \log p_k$$

subject to

1. $\forall k \in [1, l], 0 \leq p_k \leq 1,$
2. $\sum_{k=1}^{l} = 1,$ and
3. $\forall i, j: \sum_{k \in [i,j], (i,j) \in m_k} p_k = p_{i,j}.$
P-Med-Schema Consolidation

Example  $M = \{M_1, M_2\}$, where $M_1$ contains three attributes $\{a_1, a_2, a_3\}$, $\{a_4\}$, and $\{a_5, a_6\}$, and $M_2$ contains two attributes $\{a_2, a_3, a_4\}$ and $\{a_1, a_5, a_6\}$. The target schema $T$ would then contain four attributes: $\{a_1\}$, $\{a_2, a_3\}$, $\{a_4\}$, and $\{a_5, a_6\}$.

Algorithm 3: Consolidate a p-med-schema.

0: **Input**: Mediated schemas $M_1, \ldots, M_l$.
   **Output**: A consolidated single mediated schema $T$.
1: Set $T = M_1$.
2: for $(i = 2, \ldots, l)$ modify $T$ as follows:
3:   for each (attribute $A'$ in $M_i$)
4:     for each (attribute $A$ in $T$)
5:       Divide $A$ into $A \cap A'$ and $A - A'$;
6: return $T$. 

@cegme (University of Florida)
Consolidating P-Mappings

See paper ...
Experiments

- UDI Systems takes a set of data sources and creates mediated schema and probabilistic schema.
- UDI accepts select-project queries.\(^3\)
- Each data source was in a MySQL table, query processor was written in Java.
- Second String for JaroWinkler and pairwise similarity.
- Knitro to solve entropy maximization problem.
- Compared UDI with answers obtained through manual integration.

\(^3\) no joins because it is one table. They did not specify self-joins.
Goal  Compare UDI vs manual integration

Table 2: Precision, recall and F-measure of query answering of the UDI system compared with a manually created integration system. The results show that UDI obtained a high accuracy in query answering.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Golden standard</td>
<td></td>
<td></td>
</tr>
<tr>
<td>People</td>
<td>1</td>
<td>.849</td>
<td>.918</td>
</tr>
<tr>
<td>Bib</td>
<td>1</td>
<td>.852</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td>Approximate golden standard</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Movie</td>
<td>.95</td>
<td>1</td>
<td>.924</td>
</tr>
<tr>
<td>Car</td>
<td>1</td>
<td>.917</td>
<td>.957</td>
</tr>
<tr>
<td>Course</td>
<td>.958</td>
<td>.984</td>
<td>.971</td>
</tr>
<tr>
<td>People</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Bib</td>
<td>1</td>
<td>.955</td>
<td>.977</td>
</tr>
</tbody>
</table>
Competing Automatic Approaches

MySQL Keyword  Consider all data sources as text documents and apply keyword search techniques. Used MySQL’s Keyword Search Engine.

SOURCE  Answers the query on every data source with the appropriate attributes

TOPMAPPING  Consolidate schema but only take the highest probability mediated schema.
## Quality of Mediated Schema

<table>
<thead>
<tr>
<th>Domain</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie</td>
<td>.97</td>
<td>.62</td>
<td>.76</td>
</tr>
<tr>
<td>Car</td>
<td>.68</td>
<td>.88</td>
<td>.77</td>
</tr>
<tr>
<td>People</td>
<td>.76</td>
<td>.86</td>
<td>.81</td>
</tr>
<tr>
<td>Course</td>
<td>.83</td>
<td>.58</td>
<td>.68</td>
</tr>
<tr>
<td>Bib</td>
<td>.77</td>
<td>.81</td>
<td>.79</td>
</tr>
<tr>
<td>Avg</td>
<td>.802</td>
<td>.75</td>
<td>.762</td>
</tr>
</tbody>
</table>

Counted how many pairs of attributed were correctly clustered
Setup Efficiency

Setup time includes:

1. importing source schemas
2. creating p-med-schema
3. creating p-mapping between each source schema and each possible mediated schema
4. consolidating the p-med-schema and p-mappings

Figure 7: System setup time for the Car domain. When the number of data sources was increased, the setup time increased linearly.
Conclusion

- Demonstrated that it is possible to get accurate data integration working automatically
- In the future authors consider more than one table, normalization of mediated schemas, modeling data integration over time. Also, adding user feedback back into the loop (Pay-as-you-go User feedback)
Thank You!