POZNAN UNIVERSITY OF TECHNOLOGY

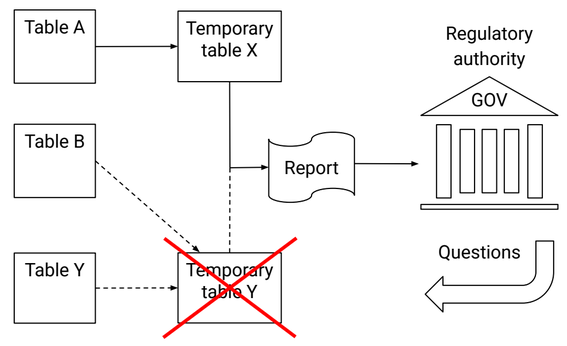
ProvenanceWeek2025@SIGMOD
June 27, 2025 @ Berlin, Germany

On fixing broken lineage

Witold Andrzejewski, Paweł Boiński, Robert Wrembel
Institute of Computing Science, Poznan University of Technology, Poland

Potential sources of broken lineage

1. The use of temporary objects in processing (e.g., tables, files) that are deleted once the task is completed
2. The use of User Defined Functions (UDFs) whose processing logic is not captured by Data Lineage systems
3. Passing the data through external systems that do not record the details of the transformation
4. Transition to new systems (migrations or upgrades) without preserving historical data about Data Lineage
5. Occasional manual data handling
6. The absence of a predefined lineage or a provenance capture mechanism



Unique key detection

1. Use metadata if available
2. Determine based on typical column names
3. Unique, *non-null* attributes:
   a. Integer numeric attributes—usually surrogate keys
   b. Text attributes—usually natural keys

...	...
100	...
101	...
102	...
...	...

Foreign key detection

1. Use metadata if available
2. Determine based on typical column names
3. Find attributes with *nulls* and non-unique values from key attributes

...	...
100	...
null	...
100	...
102	
...	...

Unique key source-target detection

1. Keys are most commonly immutable
2. Target key attributes are thus obtained via selection from source key attributes
3. Therefore, target key attribute contains a subset of source key attribute

...	...
100	...
101	...
...	...

Foreign key source-target detection

1. Keys are most commonly immutable
2. Target foreign key attributes are thus obtained via selection from source foreign key attributes
3. Therefore, target foreign key attribute contains a multi-subset of source foreign key attribute

...	...
100	...
null	...
100	...
...	...

Numeric source-target attribute detection

1. We support selection and potential linear transformation during projection
2. Z-scored values in potential source and target attributes should have the same distribution (if projection was linear transformation)
3. Perform 2-sample Kolmogorov-Smirnov test to potentially reject the possibility that one attribute was derived from the other
4. Linear transformation coefficients can be found via equations similar to a linear regression:

$$a = \frac{\sigma(Y)}{\sigma(X)} \qquad b = E(Y) - \frac{\sigma(Y)}{\sigma(X)}E(X)$$

5. Find possible source-target value pairs based on derived coefficients; take into account information about possible mapping based on derived keys
6. Preliminary experimental results (linear transformation with *a=10* and *b=50*)

		uniform			normal			uniform discreet		
$\|bag(cX)\|$	$\|bag(cY)\|$	a	b	rank	a	b	rank	a	b	rank
10000	1000	10.071	49.903	216	10.050	50.323	294	10.082	54.634	2
5000	1000	10.116	49.850	167	10.043	50.231	115	9.750	49.514	2
2000	1000	10.030	50.040	47	9.846	50.083	37	9.857	49.741	1

...	...
...	1
...	4
	7
...	10.5
...	...

$$y = ax + b$$

...	...
...	12
...	72
	107
...	...

Text source-target attribute detection

1. We support selection and potential minor mutation of the text under projection
2. Case 1, mutations include: deletions and insertions of symbols from a predefined set
   - Match rows based on inverted index built for letters in potential source attributes
3. Case 2, mutations include: insertions, modifications and deletions
   - Match rows based on the M-tree index built for text similarity metrics, e.g.: Jaccard distance, Normalized Levenshtein distance or Modified Longest Common Subsequence (MLCS)
4. Preliminary experimental results

		Substring		Levenshtein		MLCS		Jaccard	
$\|bag(cX)\|$	$\|bag(cY)\|$	Index	No Index	Index	No Index	Index	No Index	Index	No Index
10000	1000	0.386s	12.966s	2.593s	2.612s	18.523s	25.903s	6.649s	14.496s
5000	1000	0.201s	6.443s	1.318s	1.302s	9.059s	12.750s	3.316s	7.183s
2000	1000	0.080s	2.550s	0.517s	0.525s	3.603s	5.122s	1.299s	2.843s

...	...
...	John Smith
...	Mia Taylor
...	Arthur Jones
...	...

...	...
...	John T. Smythe
...	A. Jones
...	...