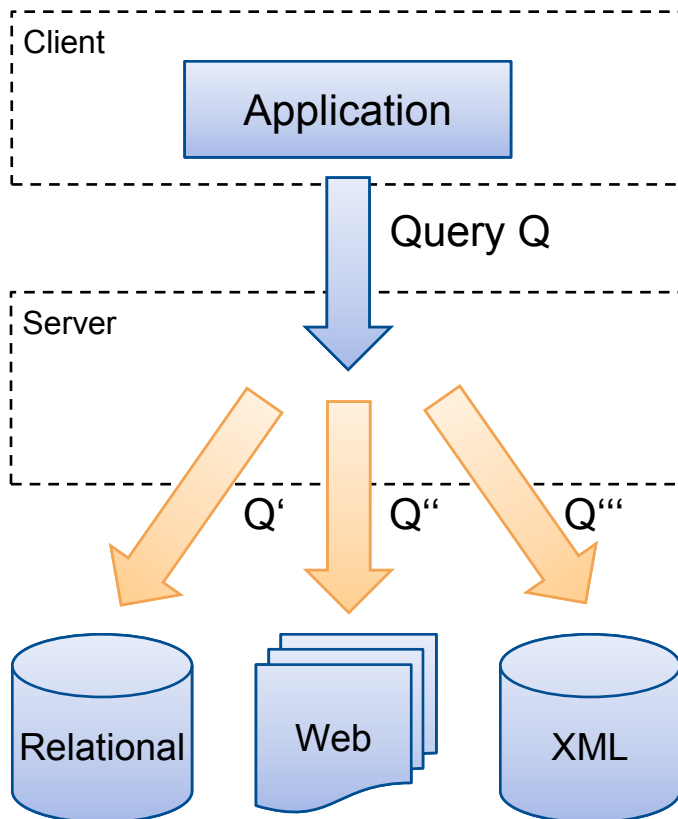


A First Step Towards Integration Independence

Laura M. Haas IBM Almaden Research
Renée J. Miller University of Toronto
Donald Kossmann ETH Zurich
Martin Hentschel ETH Zurich



Data Federation



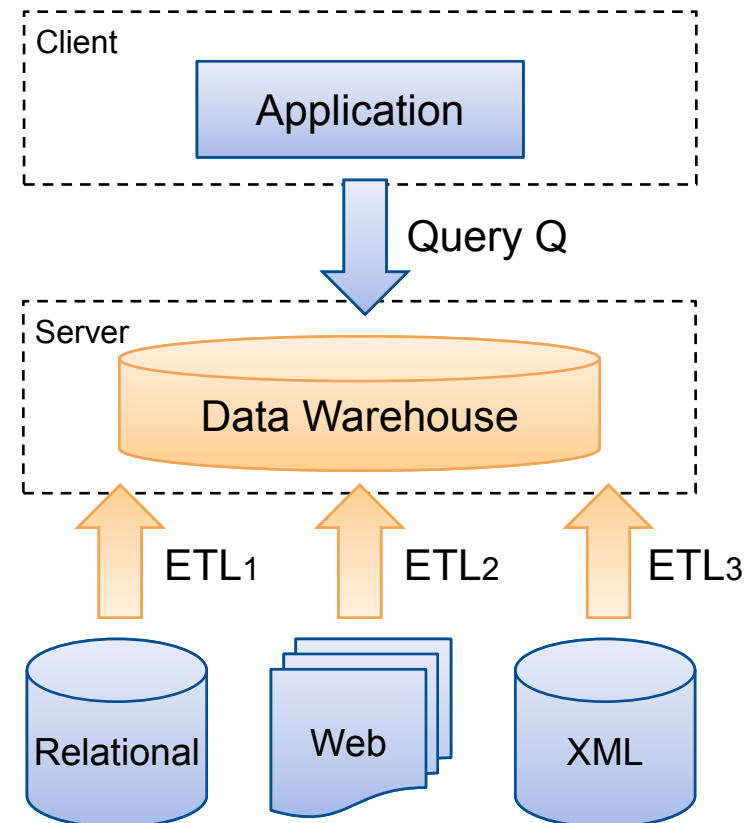
- Provides virtual, integrated view of underlying sources
- Lazy integration¹, Mediation², Virtual integration
- Garlic, Tsimmis, Information Manifold, etc.

¹ J. Widom, Integrating Heterogeneous Database: Lazy or Eager?, 1996

² G. Wiederhold, Mediators in the Architecture of Future Information Systems, 1992

Data Warehousing

- Builds a data warehouse, transforms all data
- Eager integration¹, Extract Transform Load (ETL), Materialization
- IBM InfoSphere, Informatica, Oracle, SAP, Microsoft, etc.

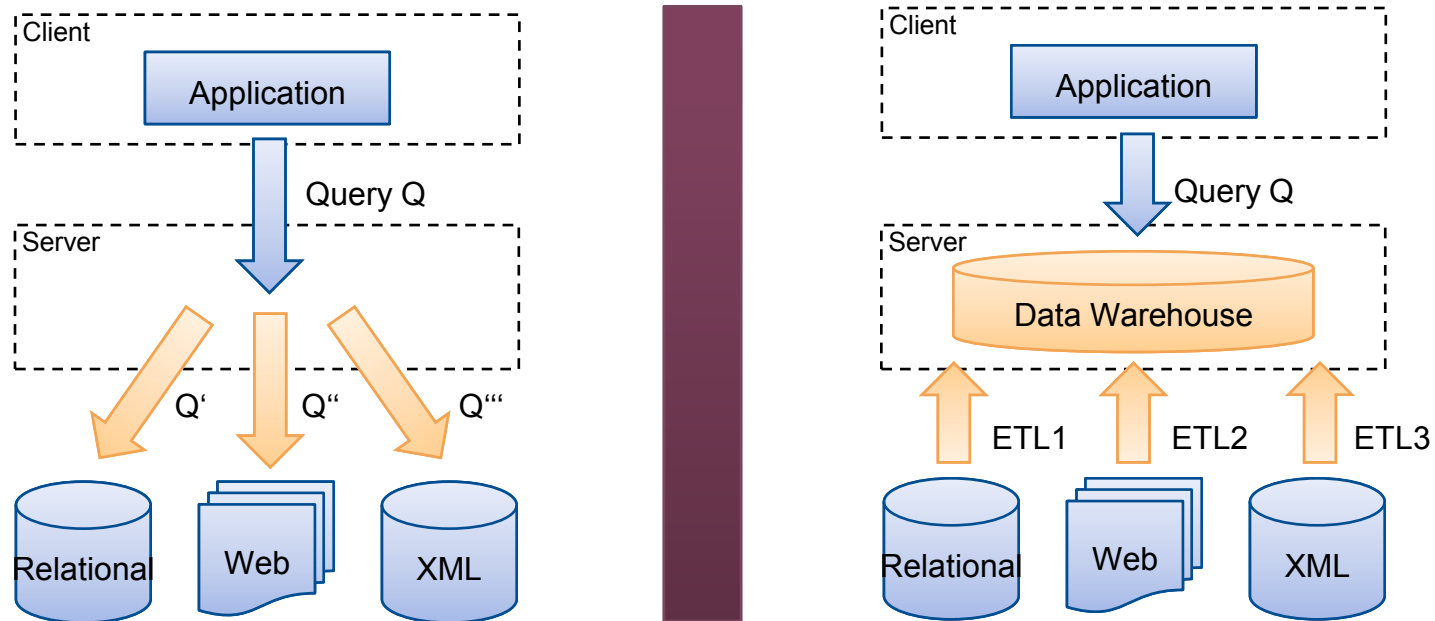


¹ J. Widom, Integrating Heterogeneous Database: Lazy or Eager?, 1996

Decision criteria

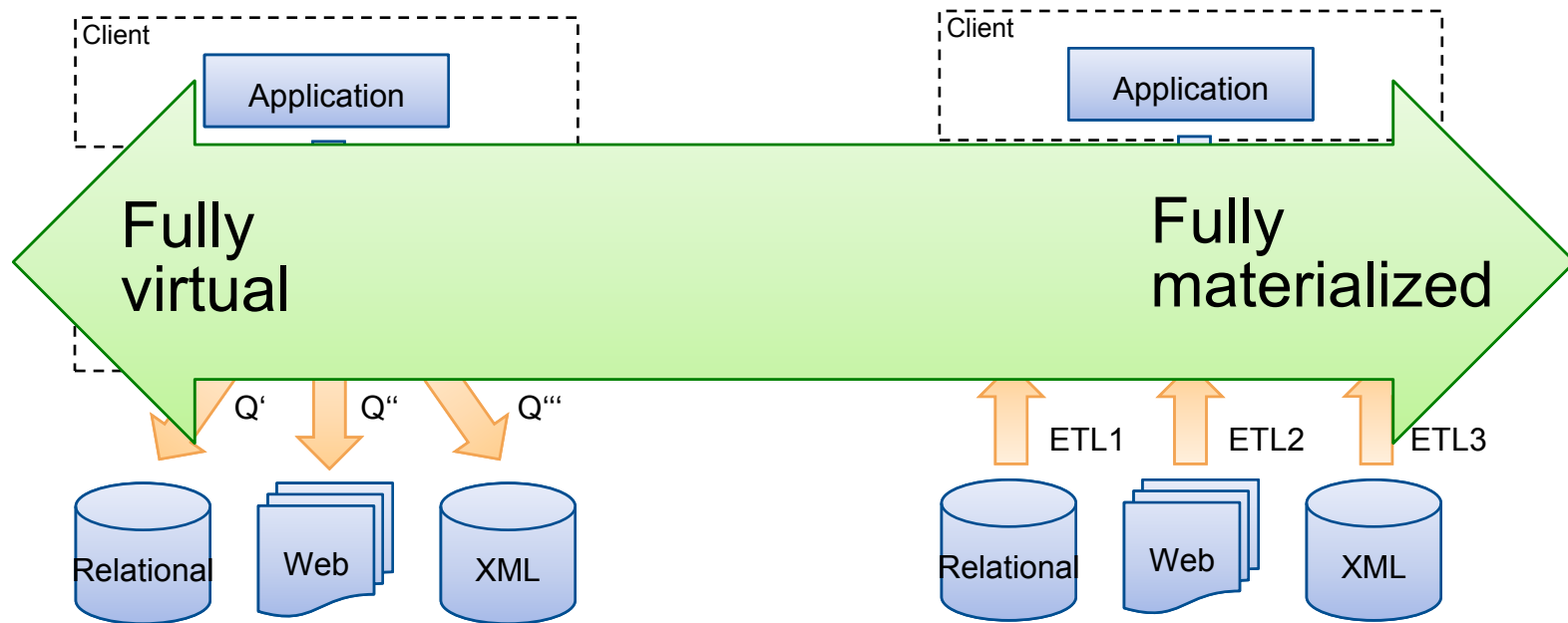
Federation	Data characteristics	Warehousing
Volatile data Small answer sets		Static data Large answer sets
Long	Response time	Short
Less critical	Availability	Critical
Fresh data	Data freshness	Stale data
Unpredictable	Workload	Predictable
Days/weeks	Time to market	Month/years
Low	Installation cost	High

Problematic gap



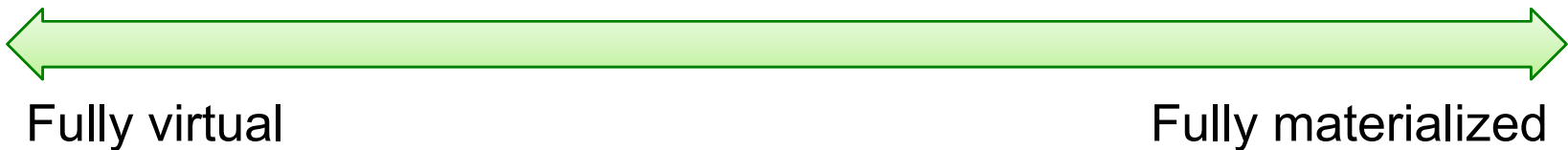
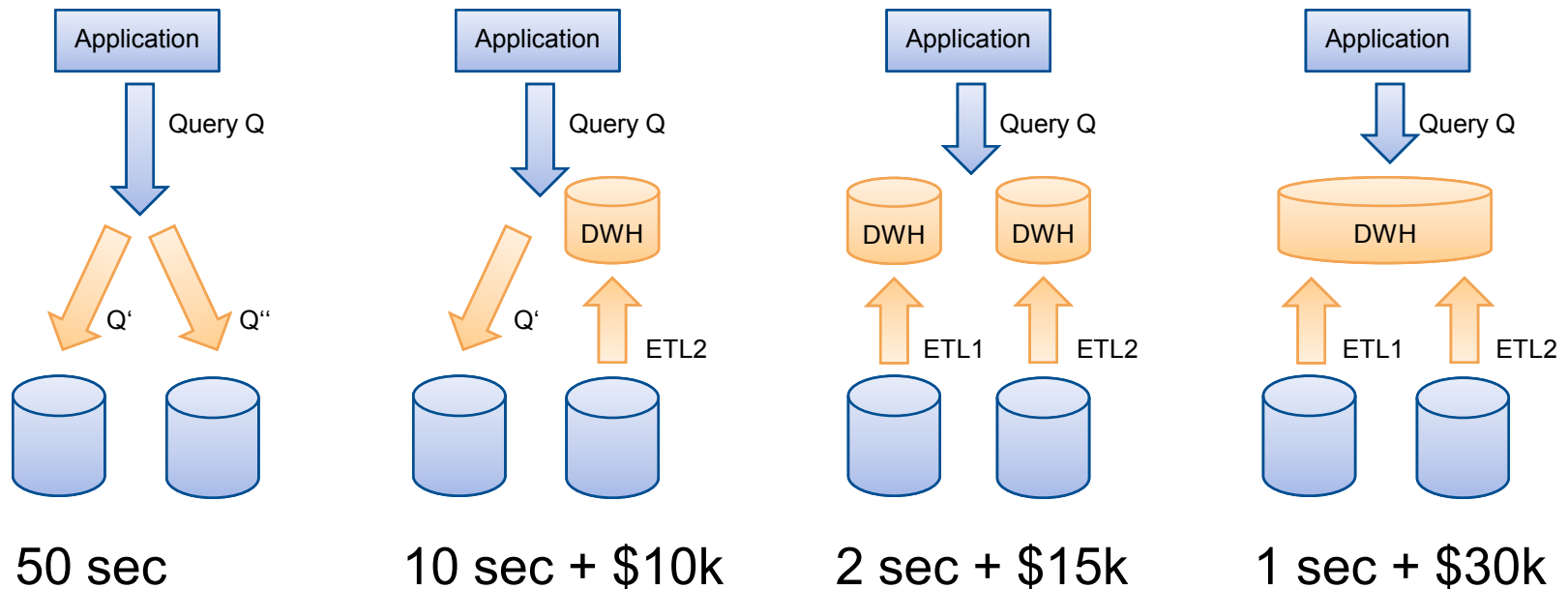
- Decision early in design process
- Different design-time tools
- Change is expensive: time and money

Integration Independence

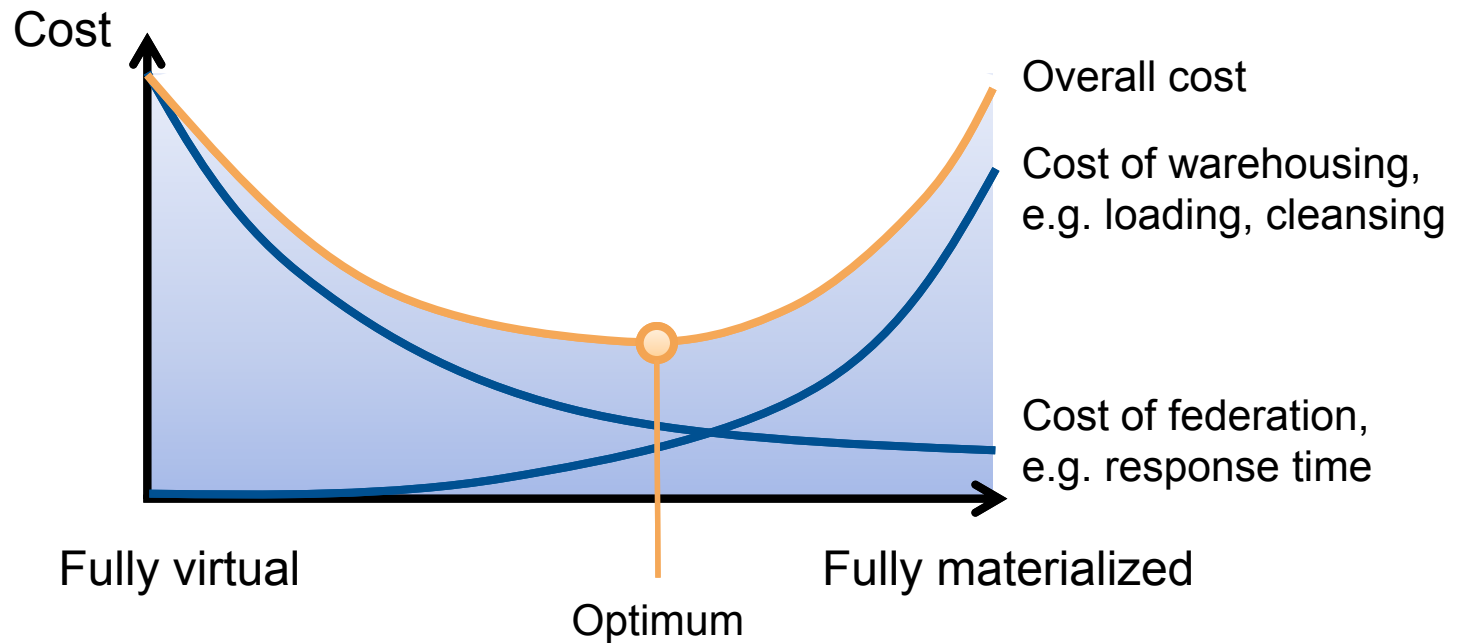


- Move between two extremes
- Changes in how, when, where to integrate transparent
- Integration becomes optimization choice

Optimizing an integration



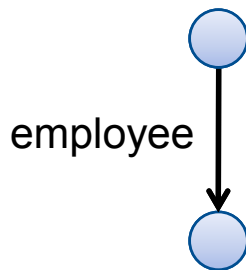
Vision



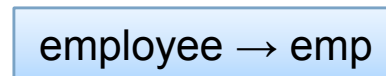
- Fine-grained movement
- Overall cost includes all decision criteria
- Automatically choose optimal integration

Mapping Data to Queries (MDQ)

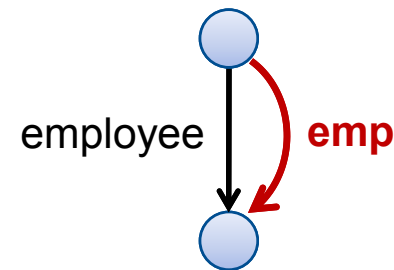
- Data integration system for XML
- Data represented as graph
- Integration through annotations



Original data



Mapping rule



Integrated data

Virtual integration with MDQ

- Sum of bonuses of emps

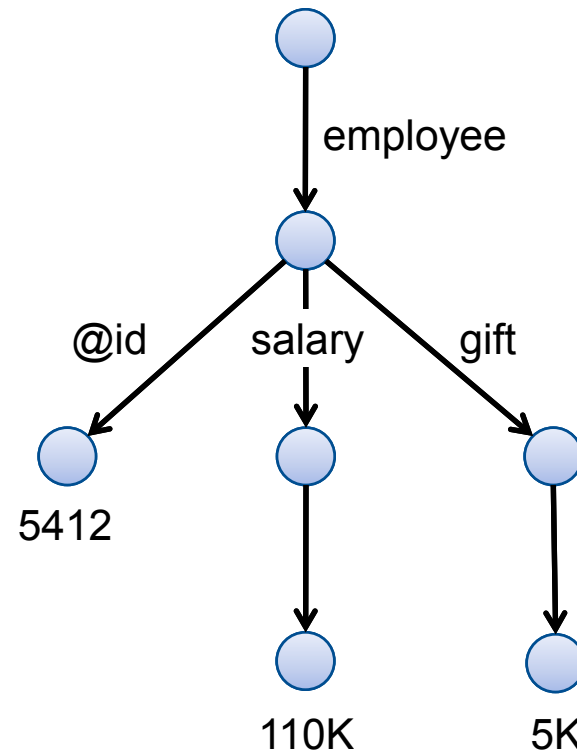
```
sum(//emp/bonus)
```

- Data

```
<employee id="5412">  
  <salary>  
    110K  
  </salary>  
  <gift>  
    5K  
  </gift>  
</employee>
```

- Mappings

```
employee → emp  
gift → bonus
```



Virtual integration with MDQ

- Sum of bonuses of emps

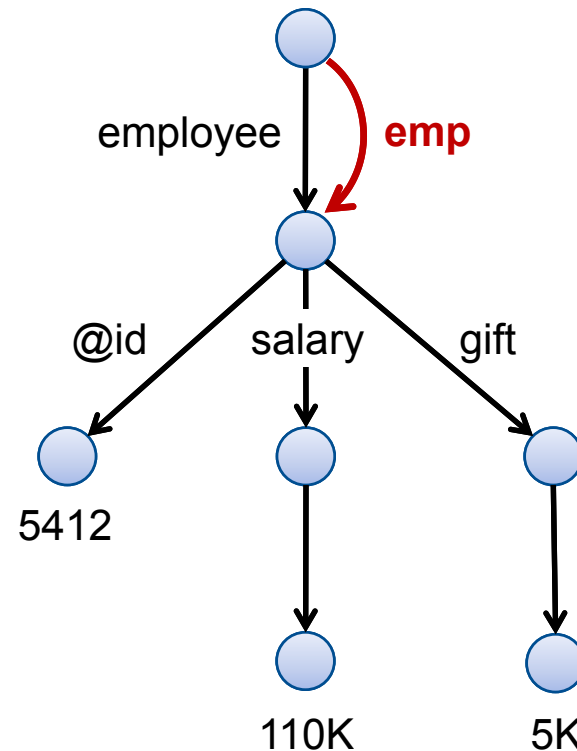
```
sum(//emp/bonus)
```

- Data

```
<employee id="5412">  
  <salary>  
    110K  
  </salary>  
  <gift>  
    5K  
  </gift>  
</employee>
```

- Mappings

```
employee → emp  
gift → bonus
```



Virtual integration with MDQ

- Sum of bonuses of emps

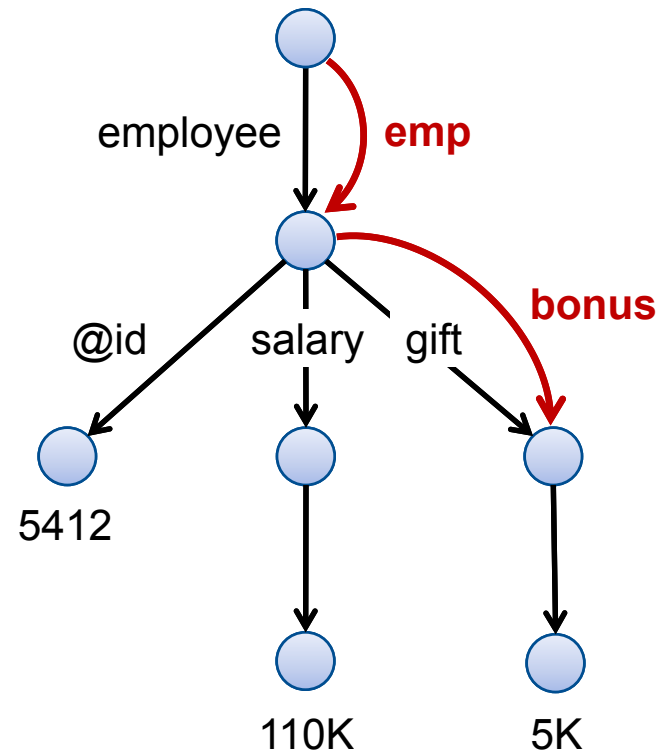
```
sum(//emp/bonus)
```

- Data

```
<employee id="5412">  
  <salary>  
    110K  
  </salary>  
  <gift>  
    5K  
  </gift>  
</employee>
```

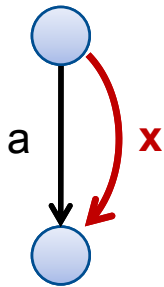
- Mappings

```
employee → emp  
gift → bonus
```

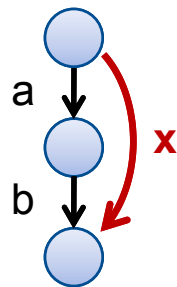


Data integration with MDQ

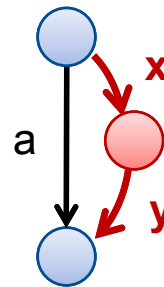
- Schema level integration¹



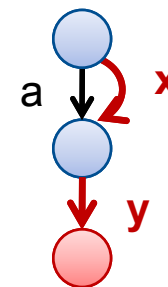
Aliasing
 $a \rightarrow x$



Un-nesting
 $a/b \rightarrow x$

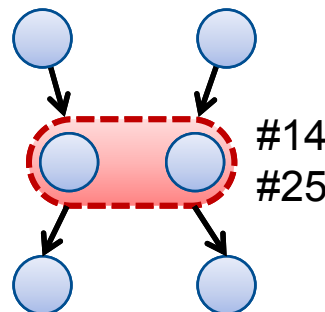


Nesting
 $a \rightarrow x/y$



Element constr.
 $a \rightarrow \langle x \rangle \langle y \rangle \langle /x \rangle$

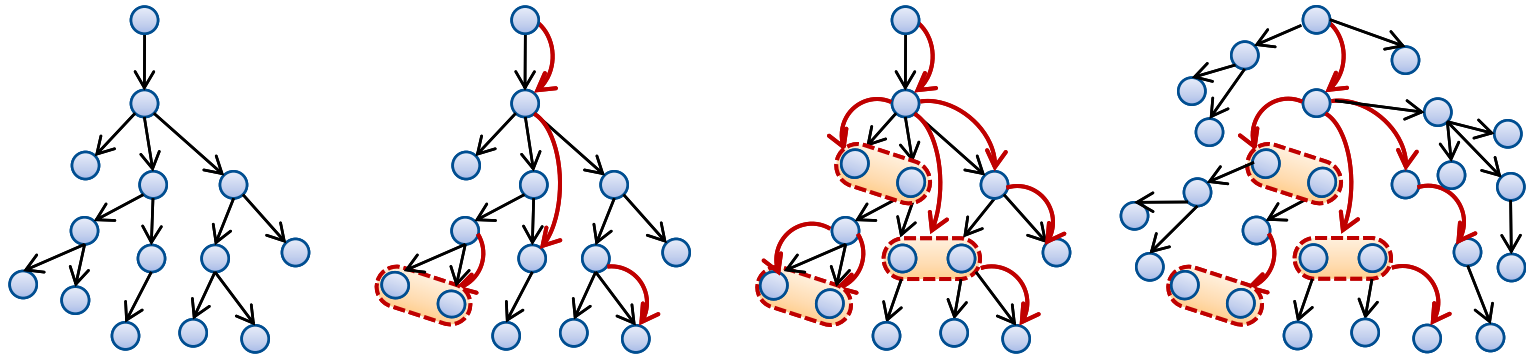
- Instance integration



De-duplication
 $\#14 \leftrightarrow \#25$

¹ Mapping scenarios in B. Alexe et al., STBenchmark, VLDB 2008

From virtual to materialized (and back)



Fully virtual

Fully materialized

- Many options to materialize annotations
e.g. parts of document, effects of mapping rule
- Cleansing, transformation, and pre-aggregation possible

Summary

- Integration Independence
 - Applications should be immune to changes in how and when information is integrated
- Vision
 - Optimize overall cost of integration
- Mapping Data to Queries
 - Graph based integration using annotations
 - Materialize annotations