CS 595 - Hot topics in database systems: **Data Provenance**

1. Introduction to Data Provenance

Boris Glavic

August 29, 2012

Outline

ILLINOIS INSTITUTE OF TECHNOLOGY

Data Provenance

Data Provenance

Information about the **creation process** and **origin** of data

ILLINOIS INSTITUTE OF TECHNOLOGY

Why do we call it Provenance?

Origin of the Term

- From art dealing

Alternative Terms

- Lineage
- Data Pedigree

ILLINOIS INSTITUTE OF TECHNOLOGY

Why do we call it Provenance?

Origin of the Term

- From art dealing

Alternative Terms

- Lineage for kings
- Data Pedigree

ILLINOIS INSTITUTE OF TECHNOLOGY

Why do we call it Provenance?

Origin of the Term

- From art dealing

Alternative Terms

- Lineage for kings
- Data Pedigree for dogs

Why do we call it Provenance?

Origin of the Term

- From art dealing for pieces of art

Alternative Terms

- Lineage for kings
- Data Pedigree for dogs

ILLINOIS INSTITUTE OF TECHNOLOGY

Provenance in Art

Given a piece of art

- How do we know ...
  - if it is authentic?
  - who created it?
  - if it has been altered?

Example

Jan Van Eyck - Arnolfini Portrait

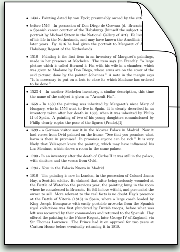
Provenance in Art

Provenance

- French *provenir*, "to come from"
- Chronology of the ownership or location of an historical object

Example

Jan Van Eyck - Arnolfini Portrait



- 1434 - Painting dated by van Eyck; presumably owned by the sitt
- before 1516 - In possession of Don Diego de Guevara (d. Brussels a Spanish career courtier of the Habsburgs (himself the subject o portrait by Michael Sittow in the National Gallery of Art). He live of his life in the Netherlands, and may have known the Arnolfinis i later years. By 1516 he had given the portrait to Margaret of A Habsburg Regent of the Netherlands.
- 1516 - Painting is the first item in an inventory of Margaret's paintings, made in her presence at Mechelen. The item says (in French): "a large picture which is called Hernoul le Fin with his wife in a chamber, which was given to Madame by Don Diego, whose arms are on the cover of the said picture; done by the painter Johannes." A note in the margin says "It is necessary to put on a lock to close it: which Madame has ordered to be done."
- 1523-4 - In another Mechelen inventory, a similar description, this time the name of the subject is given as "Arnoult Fin".
- 1558 - In 1530 the painting was inherited by Margaret's niece Mary of Hungary, who in 1556 went to live in Spain. It is clearly described in an inventory taken after her death in 1558, when it was inherited by Philip II of Spain. A painting of two of his young daughters commissioned by Philip clearly copies the pose of the figures (Prado).[1]
- 1599 - a German visitor saw it in the Alcazar Palace in Madrid. Now it had verses from Ovid painted on the frame: "See that you promise: what harm is there in promises? In promises anyone can be rich." It is very likely that Velázquez knew the painting, which may have influenced his Las Meninas, which shows a room in the same palace.
- 1700 - In an inventory after the death of Carlos II it was still in the palace, with shutters and the verses from Ovid.
- 1794 - Now in the Palacio Nuevo in Madrid.
- 1816 - The painting is now in London, in the possession of Colonel James Hay, a Scottish soldier. He claimed that after being seriously wounded at the Battle of Waterloo the previous year, the painting hung in the room where he convalesced in Brussels. He fell in love with it, and persuaded the owner to sell. More relevant to the real facts is no doubt Hay's presence at the Battle of Vitoria (1813) in Spain, where a large coach loaded by King Joseph Bonaparte with easily portable artworks from the Spanish royal collections was first plundered by British troops, before what was left was recovered by their commanders and returned to the Spanish. Hay offered the painting to the Prince Regent, later George IV of England, via Sir Thomas Lawrence. The Prince had it on approval for two years at Carlton House before eventually returning it in 1818.

1523-4	In another Mechelen inventory, a similar description, this time the name of the subject is given as "Arnoult Fin".
1558	In 1530 the painting was inherited by Margaret's niece Mary of Hungary, who in 1556 went to live in Spain. It is clearly described in an inventory taken after her death in 1558, when it was inherited by Philip II of Spain. A painting of two of his young daughters commissioned by Philip clearly copies the pose of the figures (Prado).[1]

Outline

ILLINOIS INSTITUTE OF TECHNOLOGY

Relational Algebra

- Formalizes queries over relational data
- Is an **algebra** over relations
  - Types of operators
  - An operator produces a **single** output relation from **one or more** input relations.
- **Relations**
  - A relation is a set of tuples with the same schema
  - Tuple is a list of values
- **Composable**
  - Output of an operator can be used as input to another operator!
  - $\Rightarrow$ Can build complex queries by combining simple operators

Types of Operators

- Selection σ_C
- Projection π_A
- Joins
  - Theta-join $\bowtie_C$
  - Cross-product $\times$
  - Outer joins ⟕, ⟖, ⟗
- Aggregation with group-by $\alpha_{agg,G}$
- Set Operations
  - Union $\cup$
  - Intersection $\cap$
  - Set difference $-$
- Relation access R

ILLINOIS INSTITUTE OF TECHNOLOGY

Selection

Signature

- $\sigma_C(R)$
- R is the input relation
- C is a logical condition (**selection condition**)
  - Logical operators: AND ($\wedge$), OR ($\vee$), and NOT ($\neg$)
  - Comparison operators:
    - E.g., equality ($=$) or smaller equals ($\leq$)
    - Refer to constants and attributes
  - Calls to functions?
  - E.g., $name =' Peter' \wedge salary \leq 1000$

ILLINOIS INSTITUTE OF TECHNOLOGY

Selection

Definition

$$[[\sigma_C(R)]] = \{t \mid t \in R \wedge t \models C\}$$

Selection

Example

$$Employee = \{(Peter, 100), (Heinz, 4000)\}$$
$$[[\sigma_{name='Peter'}(Employee)]] = \{(Peter, 100)\}$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Projection

Signature

- $\pi_A(R)$
- R is the input relation
- A is a list of **projection expressions**
  - Attributes form R
  - Functions calls and operators
    - E.g., $a + b$
  - Renaming, $a \rightarrow b$

Projection

Definition

$$[[\pi_A(R)]] = \{t \mid \exists u \in R \wedge u.A = t\}$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Projection

Example

$$Employee = \{(Peter, 100), (Heinz, 4000)\}$$
$$[[\pi_{salary}(Employee)]] = \{(100), (4000)\}$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Join

Signature

- $R \bowtie_C S$
- R, S are the input relations
- C is a logical condition (**join condition**)
  - Same as selection condition
  - Only equality conditions and $\wedge$ $\Rightarrow$Equi-join

ILLINOIS INSTITUTE OF TECHNOLOGY

Join

Definition

$$[[R \bowtie_C S]] = \{t \blacktriangleright t' \mid t \in R \wedge t' \in S \wedge t \blacktriangleright t' \models C\}$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Join

Example

$$Employee = \{(Peter, 1), (Heinz, 2)\}$$
$$Department = \{(1, CS), (2, HR)\}$$
$$[[Employee \bowtie_{depId=Id} Department]] = \{(Peter, 1, 1, CS), (Heinz, 2, 2, HR)\}$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Outer Joins: Left-outer Join

Signature

- $R ⟕_C S$
- R, S are the input relations
- C is a logical condition (**join condition**)

ILLINOIS INSTITUTE OF TECHNOLOGY

Outer Joins: Left-outer Join

Definition

$$[[R ⟕_C S]] = \{(t \blacktriangleright t') \mid t \in R \wedge t' \in S\}$$
$$\cup \{(t_1 \blacktriangleright null(S)) \mid t \in R \wedge (\nexists t' \in S : (t \blacktriangleright t') \models C)\}$$

Outer Joins: Left-outer Join

Example

$$
\begin{aligned}
Employee &= \{(Peter, 1), (Heinz, null)\} \\
Department &= \{(1, CS), (2, HR)\} \\
[[Employee ⟕_{depId=Id} Department]] &= \{(Peter, 1, 1, CS), \\
&\quad (Heinz, null, null, null)\}
\end{aligned}
$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Aggregation

Signature

- $\alpha_{agg,G}(R)$
- R is the input relation
- *agg* list of aggregation functions
  - E.g., $sum(a)$ if a attribute of R
- G is a list of group-by expressions
  - Attributes
  - Operators and function expressions

ILLINOIS INSTITUTE OF TECHNOLOGY

Aggregation

Definition

$$[[\alpha_{G,agg}(R)]] = \{(t.G, res_1, \ldots, res_m) \mid t \in R \\ \wedge \forall i \in \{1, m\} : res_i = agg_i(\pi_{b_i}(\sigma_{G=t.G}(R))\}$$

- b_i expression used as aggregation function input
  - E.g., a for $sum(a)$
- res_i is result of computing aggregation function for a tuple

ILLINOIS INSTITUTE OF TECHNOLOGY

Aggregation

Example

$$Employee = \{(Peter, 1, 3000), (Heinz, 2, 4000), (Jule, 1, 2000)\}$$

$$[[\alpha_{sum(salary), depId}]] = \{(5000, 1), (4000, 2)\}$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Union

Signature

- $R \cup S$
- R and S are the input relations
- R and S have to have same schema

Union

Definition

$$[[R \cup S]] = \{t \mid t \in R \vee t \in S\}$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Union

Example

$$Employee = \{(Peter), (Heinz), (Jule)\}$$
$$Manager = \{(Peter), (Gertrud)\}$$
$$[[Employee \cup Manager]] = \{(Peter), (Heinz), (Jule), (Getrud)\}$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Intersection

Signature

- $R \cap S$
- R and S are the input relations
- R and S have to have same schema

ILLINOIS INSTITUTE OF TECHNOLOGY

Intersection

Definition

$$[[R \cap S]] = \{t \mid t \in R \land t \in S\}$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Intersection

Example

$$\begin{aligned} Employee &= \{(Peter), (Heinz), (Jule)\} \\ Manager &= \{(Peter), (Gertrud)\} \\ [[Employee \cap Manager]] &= \{(Peter)\} \end{aligned}$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Set Difference

Signature

- $R - S$
- R and S are the input relations
- R and S have to have same schema

ILLINOIS INSTITUTE OF TECHNOLOGY

Set Difference

Definition

$$[[R - S]] = \{t \mid t \in R \wedge t \notin S\}$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Set Difference

Example

$$Employee = \{(Peter), (Heinz), (Jule)\}$$
$$Manager = \{(Peter), (Gertrud)\}$$
$$[[Employee - Manager]] = \{(Heinz), (Jule)\}$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag vs. Set semantics

- So far: Relations are sets (Set semantics)
  - ⇒A tuple appears at most one time

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag vs. Set semantics

- So far: Relations are sets (Set semantics)
  - ⇒A tuple appears at most one time
- This is different from SQL and database implementations
  - Tuple can appear more then once

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag vs. Set semantics

- So far: Relations are sets (Set semantics)
  - ⇒A tuple appears at most one time
- This is different from SQL and database implementations
  - Tuple can appear more then once
  - In relation in DB only if no Primary key

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag vs. Set semantics

- So far: Relations are sets (Set semantics)
  - ⇒A tuple appears at most one time
- This is different from SQL and database implementations
  - Tuple can appear more then once
  - In relation in DB only if no Primary key
  - This is called Bag semantics

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag vs. Set semantics

- So far: Relations are sets (Set semantics)
  - ⇒A tuple appears at most one time
- This is different from SQL and database implementations
  - Tuple can appear more then once
  - In relation in DB only if no Primary key
  - This is called Bag semantics
- Bag semantics
  - Formally: assign a multiplicity ≥ 1 to each tuple in a relation

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag vs. Set semantics cont.

- Why set semantics?
  - Cleaner formalism

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag vs. Set semantics cont.

- Why set semantics?
  - Cleaner formalism
- Why bag semantics?

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag vs. Set semantics cont.

- Why set semantics?
  - Cleaner formalism
- Why bag semantics?
  - Correctness

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag vs. Set semantics cont.

- Why set semantics?
  - Cleaner formalism
- Why bag semantics?
  - Correctness
    - (e.g., projecting on non-unique attribute, then aggregate)

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag vs. Set semantics cont.

- Why set semantics?
  - Cleaner formalism
- Why bag semantics?
  - Correctness
    - (e.g., projecting on non-unique attribute, then aggregate)
  - Performance

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag vs. Set semantics cont.

- Why set semantics?
  - Cleaner formalism
- Why bag semantics?
  - Correctness
    - (e.g., projecting on non-unique attribute, then aggregate)
  - Performance
    - Some operators require costly duplicate removal under set semantics

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag semantics: Notation

How to write multiplicities

- Use power notation to express the multiplicity of a tuple
  - $t^n \in R$ denotes tuple t exists with multiplicity n in relation R

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag semantics: Operators

Duplicate Removal

- $\delta(R)$
- Returns a copy of R with all multiplicities set to one

ILLINOIS INSTITUTE OF TECHNOLOGY

Bag semantics: Other operators

Definitions

$$
\begin{aligned}
[[\pi_A(R)]] =& \{t^n \mid n = \sum_{u^m \in R \wedge u.A = t} m\} \\
[[\sigma_C(R)]] =& \{t^n \mid t^n \in R \wedge t \models C\} \\
[[\alpha_{G,agg}(R)]] =& \{(t.G, res_1, \ldots, res_m)^1 \mid t^n \in R \\
& \wedge \forall i \in \{1, m\} : res_i = agg_i(\pi_{B_i}(\sigma_{G=t.G}(R))\} \\
[[R \bowtie_C S)]] =& \{(t_1 \blacktriangleright t_2)^{n \times m} \mid {t_1}^n \in R \wedge {t_2}^m \in S \\
& \wedge (t_1 \blacktriangleright t_2) \models C\} \\
[[R \mathbin{=\!\!\bowtie}_C S]] =& \{(t_1 \blacktriangleright t_2)^{n \times m} \mid {t_1}^n \in R \wedge {t_2}^m \in S\} \\
& \cup \{(t_1 \blacktriangleright null(S))^n \mid {t_1}^n \in R \\
& \wedge (\nexists t_2 \in S : (t_1 \blacktriangleright t_2) \models C)\}
\end{aligned}
$$

Bag semantics: Other operators

Definitions

$$[[R \cup S]] = \{t^{n+m} \mid t^n \in R \wedge t^m \in S\}$$
$$[[R \cap S]] = \{t^{min(n,m)} \mid t^n \in R \wedge t^m \in S\}$$
$$[[R - S]] = \{t^{n-m} \mid t^n \in R \wedge t^m \in S\}$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Outline

Data Provenance

Data Provenance

Information about the **creation process** and **origin** of data

ILLINOIS INSTITUTE OF TECHNOLOGY

Provenance in Data Processing

Given a piece of data

- How do we know ...
  - which data it is derived from?
  - which transformations (SQL) where used to create it?
  - who created it?
  - ...

Example

result

	shop	rev
t_1	Migros	125
t_2	Coop	25

ILLINOIS INSTITUTE OF TECHNOLOGY

Provenance in Data Processing

Given a piece of data

- How do we know ...
  - which data it is derived from?
  - which transformations (SQL) where used to create it?
  - who created it?
  - ...

Example

Compute the revenue for each shop as sum of prices of items sold

Example

result

	shop	rev
t_1	Migros	125
t_2	Coop	25

↑

```sql
SELECT shop,
    sum(price) AS rev
FROM sales, items
WHERE itemId = id
GROUP BY shop
```

↑ ↑

sales

	shop	itemId
s_1	Migros	1
s_2	Migros	3
s_3	Coop	3

items

	id	price
i_1	1	100
i_2	2	10
i_3	3	25

Provenance in Data Processing

Given a piece of data

- How do we know ...
  - which data it is derived from?
  - which transformations (SQL) where used to create it?
  - who created it?
  - ...

Definition (Data Provenance)

Information about the origin and creation process of data.

Example

result

	shop	rev
t_1	Migros	125
t_2	Coop	25

↑

```
SELECT shop,
    sum(price) AS rev
FROM sales, items
WHERE itemId = id
GROUP BY shop
```

↑ ↑

sales

	shop	itemId
s_1	Migros	1
s_2	Migros	3
s_3	Coop	3

items

	id	price
i_1	1	100
i_2	2	10
i_3	3	25

Abstract View

Data

- Structured? Schemata?
- Atomic units? (Data items)

ILLINOIS INSTITUTE OF TECHNOLOGY

Abstract View

Data

- Structured? Schemata?
- Atomic units? (Data items)

Transformations

- Consume input data
- Produce output data
- Hierarchical composition?
- Fixed set of atomic operations?

ILLINOIS INSTITUTE OF TECHNOLOGY

Abstract View

Data

- Structured? Schemata?
- Atomic units? (Data items)

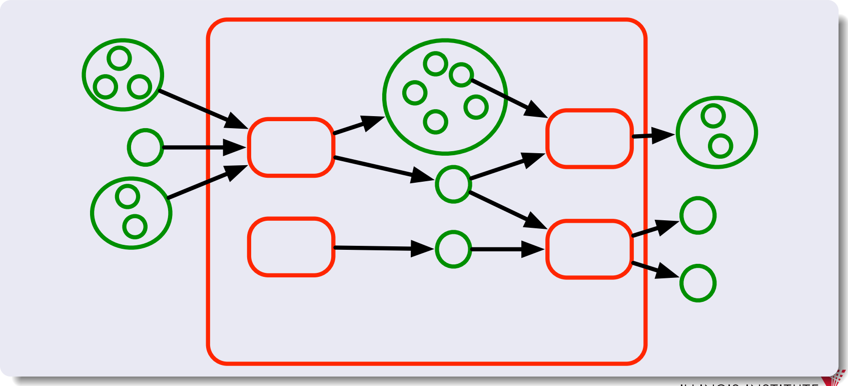
Transformations

- Consume input data
- Produce output data
- Hierarchical composition?
- Fixed set of atomic operations?

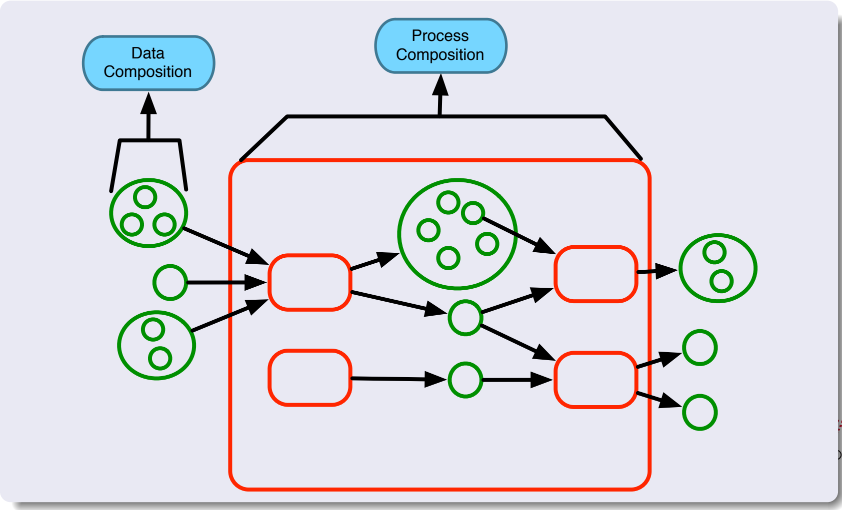
Provenance

Information about the creation process and origin of data

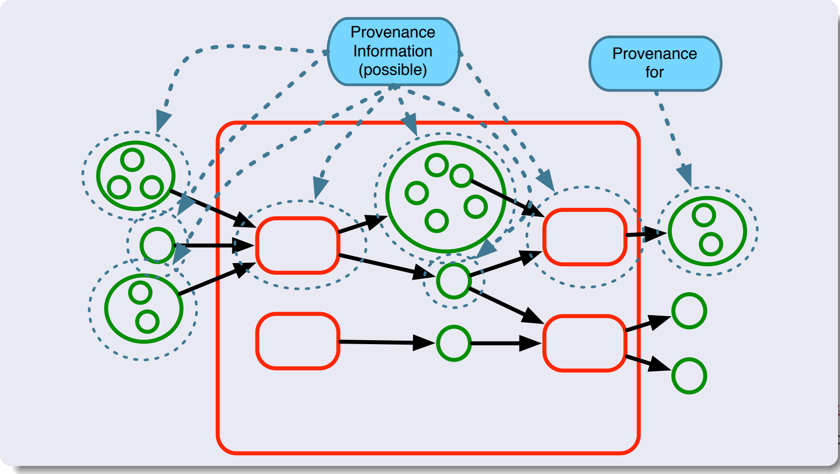
Abstract View



Abstract View



Abstract View



Scenario

- You are an analyst for a garden supply shop
- You have to compute the first quater revenue for each shop location
- Datawarehouse with sales data
- Use SQL to compute the required information from the warehouse

Running Example

Example (Input Data)

Employee

SSN	Name	WorksFor
123	Peter Peterson	New York
342	Jane Janeson	New York
555	Heinz Heinzmann	Wuppertal

Shop

Location	Budget
New York	1.000.000
Wuppertal	4.000

Item

Id	Description	Price
1	Lawnmower	199
2	Fertilizer	32
3	Rake	9

Sales

Employee	Item	Amount	Month
123	1	1	1
342	2	64	1
342	3	2	3
555	3	1	5

OF TECHNOLOGY

Running Example

Example (SalesTotal Query)

```
CREATE VIEW SalesTotal AS
SELECT Location AS Shop, Month, SSN AS Employee,
       Price * Amount AS Totalprice
FROM Employee E, Shop H, Item I, Sales S
WHERE E.WorksFor = H.Location
      AND E.SSN = S.Employee
      AND I.Id = S.Item
```

Example (Results)

SalesTotal

Shop	Month	Employee	Totalprice
New York	1	123	199
New York	1	342	2048
New York	3	342	18
Wuppertal	5	555	9

Running Example

Example (MonthlyRevenue Query)

```
CREATE VIEW MonthlyRevenue
SELECT Shop, Month, sum(Totalprice) AS Revenue
FROM SalesTotal
GROUP BY Shop, Month
```

Example (Results)

MonthlyRevenue

Shop	Month	Revenue
New York	1	2247
New York	3	18
Wuppertal	5	9

Running Example

Example (RevenueFirstQ Query)

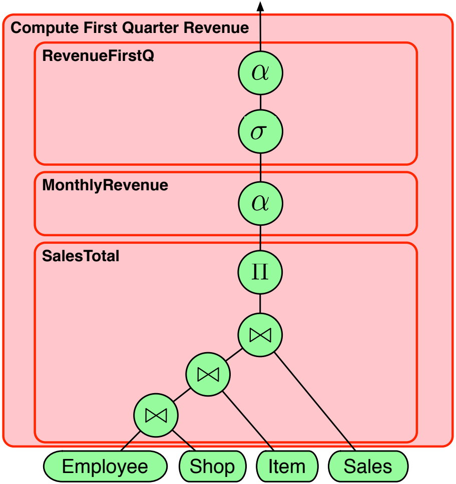
```
CREATE VIEW RevenueFirstQ
SELECT Shop, sum(Revenue) AS Revenue
FROM MonthlyRevenue
WHERE Month < 5
GROUP BY Shop
```

Example (Results)

RevenueFirstQ

Shop	Revenue
New York	2265

Running Example



Tracing an Error

Problem

- One result tuple of your query looks suspicious
- You expect the input data to be the culprit
- How to know which input data affected which output data

ILLINOIS INSTITUTE OF TECHNOLOGY

Tracing an Error

Problem

- One result tuple of your query looks suspicious
- You expect the input data to be the culprit
- How to know which input data affected which output data

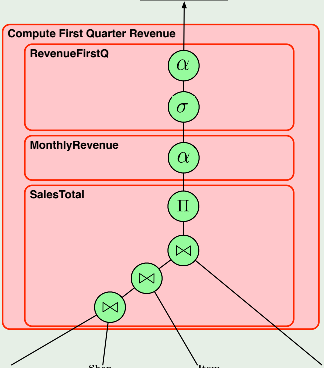
This is Data Provenance

Example Data

Example

RevenueFirstQ

Shop	Revenue
New York	2265



Employee

SSN	Name	WorksFor
123	Peter Peterson	New York
342	Jane Janeson	New York
555	Heinz Heinzmann	Wuppertal

Shop

Location	Budget
New York	1.000.000
Wuppertal	4.000

Item

Id	Description	Price
1	Lawnmower	199
2	Fertilizer	32
3	Rake	9

Sales

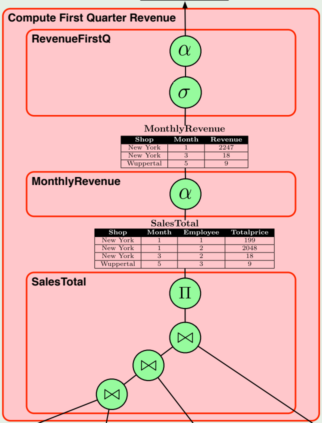
Employee	Item	Amount	Month
1	1	1	1
2	2	64	1
2	3	2	3
3	3	1	5

Example Data

Example

RevenueFirstQ

Shop	Revenue
New York	2265



MonthlyRevenue

Shop	Month	Revenue
New York	1	2247
New York	3	18
Wuppertal	5	9

SalesTotal

Shop	Month	Employee	Totalprice
New York	1	1	199
New York	1	2	2048
New York	3	2	18
Wuppertal	5	3	9

Employee

SSN	Name	WorksFor
123	Peter Peterson	New York
342	Jane Janeson	New York

Shop

Location	Budget
New York	1.000.000
Wuppertal	4.000

Item

Id	Description	Price
1	Lawnmower	199
2	Fertilizer	32

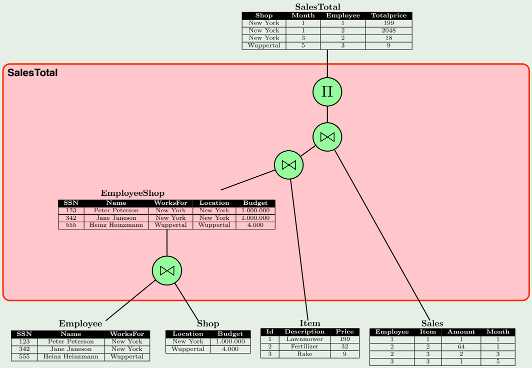
Sales

Employee	Item	Amount	Month
1	1	1	1
2	2	64	1

ILLINOIS INSTITUTE OF TECHNOLOGY

Example Data

Example



SalesTotal

Shop	Month	Employee	Totalprice
New York	1	1	199
New York	1	2	2048
New York	3	2	18
Wuppertal	5	3	9

EmployeeShop

SSN	Name	WorksFor	Location	Budget
123	Peter Peterson	New York	New York	1.000.000
342	Jane Janeson	New York	New York	1.000.000
555	Heinz Heinzmann	Wuppertal	Wuppertal	4.000

Employee

SSN	Name	WorksFor
123	Peter Peterson	New York
342	Jane Janeson	New York
555	Heinz Heinzmann	Wuppertal

Shop

Location	Budget
New York	1.000.000
Wuppertal	4.000

Item

Id	Description	Price
1	Lawnmower	199
2	Fertilizer	32
3	Rake	9

Sales

Employee	Item	Amount	Month
1	1	1	1
2	2	64	1
2	3	2	3
3	3	1	5

ILLINOIS INSTITUTE OF TECHNOLOGY

Lessons learned

Lessons learned

- Which inputs belong to provenance of outputs?

ILLINOIS INSTITUTE OF TECHNOLOGY

Lessons learned

- Which inputs belong to provenance of outputs?
  - hard

ILLINOIS INSTITUTE OF TECHNOLOGY

Lessons learned

- Which inputs belong to provenance of outputs?
  - hard
- Even if we know: How to get it?

ILLINOIS INSTITUTE OF TECHNOLOGY

Lessons learned

- Which inputs belong to provenance of outputs?
  - hard
- Even if we know: How to get it?
- Manually?

ILLINOIS INSTITUTE OF TECHNOLOGY

Lessons learned

- Which inputs belong to provenance of outputs?
  - hard
- Even if we know: How to get it?
- Manually?
  - Not reasonable for big data or complex query!

ILLINOIS INSTITUTE OF TECHNOLOGY

Lessons learned

- Which inputs belong to provenance of outputs?
  - hard
- Even if we know: How to get it?
- Manually?
  - Not reasonable for big data or complex query!
- Need system that tracks it automatically!

Outline

Types of Provenance Information

Provenance Types

- Data Provenance
- Transformation Provenance
- Additional Information

ILLINOIS INSTITUTE OF TECHNOLOGY

Types of Provenance Information

Provenance Types

- Data Provenance
  - From which input data is which output data derived from
- Transformation Provenance
- Additional Information

ILLINOIS INSTITUTE OF TECHNOLOGY

Types of Provenance Information

Provenance Types

- Data Provenance
  - From which input data is which output data derived from
- Transformation Provenance
  - Which transformations contributed in which way to which output data
- Additional Information

ILLINOIS INSTITUTE OF TECHNOLOGY

Types of Provenance Information

Provenance Types

- Data Provenance
  - From which input data is which output data derived from
- Transformation Provenance
  - Which transformations contributed in which way to which output data
- Additional Information
  - Execution environment (state of the world)
  - Involved Users

ILLINOIS INSTITUTE OF TECHNOLOGY

Data Provenance

Which/How input data influences output data

Data Provenance

Which/How input data influences output data

- Data Granularity
  - Attribute value
  - Tuple
  - Relation

Data Provenance

Which/How input data influences output data

- Data Granularity
  - Attribute value
  - Tuple
  - Relation
- Transformation Granularity
  - Query with view unfolding
  - Query block
  - Algebra operator

Data Provenance

Which/How input data influences output data

- Data Granularity
  - Attribute value
  - Tuple
  - Relation
- Transformation Granularity
  - Query with view unfolding
  - Query block
  - Algebra operator
- "True" Data Dependencies?
  - Black-box: An output depends on all inputs
  - Fine-grained: Dependencies depending on how data is processed by transformation

Data Granularity

Example (Relation)

RevenueFirstQ

Shop	Revenue
New York	2265

RevenueFirstQ
```
CREATE VIEW RevenueFirstQ
SELECT Shop, sum(Revenue) AS Revenue
FROM MonthlyRevenue
WHERE Month < 5
GROUP BY Shop
```

MonthlyRevenue

Shop	Month	Revenue
New York	1	2247
New York	3	18
Wuppertal	5	9

OF TECHNOLOGY

Data Granularity

Example (Tuple)

RevenueFirstQ

Shop	Revenue
New York	2265

RevenueFirstQ

```
CREATE VIEW RevenueFirstQ
SELECT Shop, sum(Revenue) AS Revenue
FROM MonthlyRevenue
WHERE Month < 5
GROUP BY Shop
```

MonthlyRevenue

Shop	Month	Revenue
New York	1	2247
New York	3	18
Wuppertal	5	9

Data Granularity

Example (Attribute Value)

RevenueFirstQ

Shop	Revenue
New York	2265

RevenueFirstQ

```sql
CREATE VIEW RevenueFirstQ
SELECT Shop, sum(Revenue) AS Revenue
FROM MonthlyRevenue
WHERE Month < 5
GROUP BY Shop
```

MonthlyRevenue

Shop	Month	Revenue
New York	1	2247
New York	3	18
Wuppertal	5	9

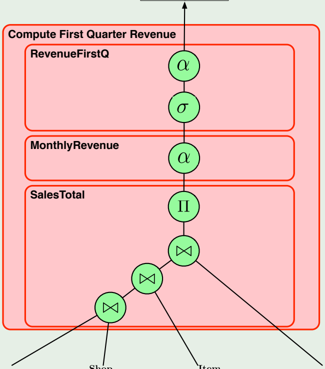
OF TECHNOLOGY

Transformation Granularity

Example

RevenueFirstQ

Shop	Revenue
New York	2265



Employee

SSN	Name	WorksFor
123	Peter Peterson	New York
342	Jane Janeson	New York
555	Heinz Heinzmann	Wuppertal

Shop

Location	Budget
New York	1.000.000
Wuppertal	4.000

Item

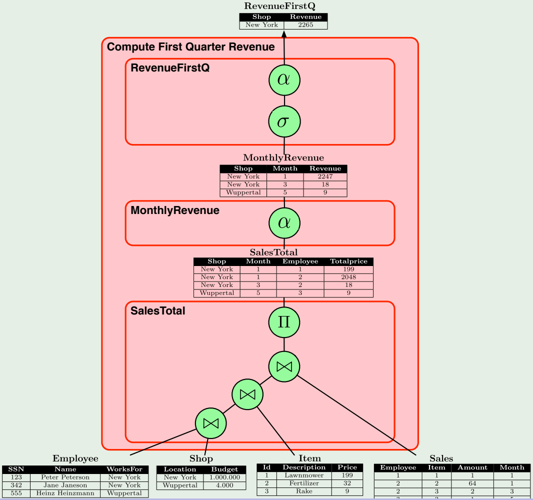
Id	Description	Price
1	Lawnmower	199
2	Fertilizer	32
3	Rake	9

Sales

Employee	Item	Amount	Month
1	1	1	1
2	2	64	1
2	3	2	3
3	3	1	5

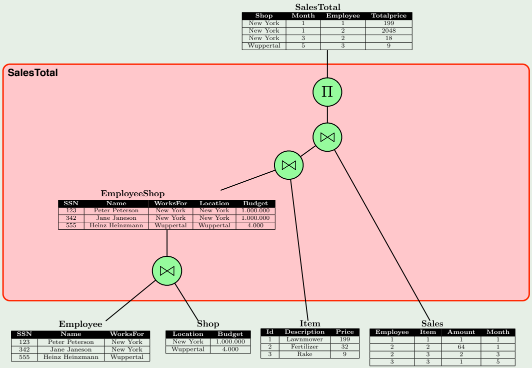
Transformation Granularity

Example



Transformation Granularity

Example



SalesTotal

Shop	Month	Employee	Totalprice
New York	1	1	199
New York	1	2	2048
New York	3	2	18
Wuppertal	5	3	9

EmployeeShop

SSN	Name	WorksFor	Location	Budget
123	Peter Peterson	New York	New York	1.000.000
342	Jane Janeson	New York	New York	1.000.000
555	Heinz Heinzmann	Wuppertal	Wuppertal	4.000

Employee

SSN	Name	WorksFor
123	Peter Peterson	New York
342	Jane Janeson	New York
555	Heinz Heinzmann	Wuppertal

Shop

Location	Budget
New York	1.000.000
Wuppertal	4.000

Item

Id	Description	Price
1	Lawnmower	199
2	Fertilizer	32
3	Rake	9

Sales

Employee	Item	Amount	Month
1	1	1	1
2	2	64	1
2	3	2	3
3	3	1	5

Data Dependencies

Example (Black-box)

RevenueFirstQ

Shop	Revenue
New York	2265

RevenueFirstQ

```
CREATE VIEW RevenueFirstQ
SELECT Shop, sum(Revenue) AS Revenue
FROM MonthlyRevenue
WHERE Month < 5
GROUP BY Shop
```

MonthlyRevenue

Shop	Month	Revenue
New York	1	2247
New York	3	18
Wuppertal	5	9

OF TECHNOLOGY

Data Dependencies

Example (Fine-grained)

RevenueFirstQ

Shop	Revenue
New York	2265

RevenueFirstQ

```sql
CREATE VIEW RevenueFirstQ
SELECT Shop, sum(Revenue) AS Revenue
FROM MonthlyRevenue
WHERE Month < 5
GROUP BY Shop
```

MonthlyRevenue

Shop	Month	Revenue
New York	1	2247
New York	3	18
Wuppertal	5	9

OF TECHNOLOGY

Transformation Provenance

Which/How transformations contributed to output data

- Transformations that generated output (transitive?)
- Only the ones that had actual effect
- Workflow template/program vs. workflow run/execution

ILLINOIS INSTITUTE OF TECHNOLOGY

Additional Information

A small subset

- OS version

Additional Information

A small subset

- OS version
- Version of library linked against

Additional Information

A small subset

- OS version
- Version of library linked against
- Environment variables

Additional Information

A small subset

- OS version
- Version of library linked against
- Environment variables
- User that executed the process

Additional Information

A small subset

- OS version
- Version of library linked against
- Environment variables
- User that executed the process
- Current main memory content

Additional Information

A small subset

- OS version
- Version of library linked against
- Environment variables
- User that executed the process
- Current main memory content
- Room temperature?

Additional Information

A small subset

- OS version
- Version of library linked against
- Environment variables
- User that executed the process
- Current main memory content
- Room temperature?
- Geographical location

Additional Information

A small subset

- OS version
- Version of library linked against
- Environment variables
- User that executed the process
- Current main memory content
- Room temperature?
- Geographical location
- ...

Additional Information

A small subset

- OS version
- Version of library linked against
- Environment variables
- User that executed the process
- Current main memory content
- Room temperature?
- Geographical location
- . . .
- Butterfly that flapped in china

Outline

ILLINOIS INSTITUTE OF TECHNOLOGY

Use Cases

- Debugging (tracking the sources of errors)
- Propagating annotations
- Gain deeper understanding of data and transformations
  - Estimate quality, trust
- Improvement of other data processing technologies
  - Probabilistic databases
  - Deletion propagation
  - Testing

ILLINOIS INSTITUTE OF TECHNOLOGY

Application Domains

- Complex database queries, e.g., datawarehousing
- E-science and curated databases
- Data integration/exchange
- Workflow systems

Application Domains

- Complex database queries, e.g., datawarehousing
- E-science and curated databases
- Data integration/exchange
- Workflow systems
- $\Rightarrow$ Application domain with complex, multi-stage data processing
  - Map-Reduce style processing and its “frontends” like Pig
  - Simulations
  - ...

Debugging

Origin of Result Tuples

- Tuple in result suspicious/wrong/interesting

Example

SalesTotal

Shop	Month	Employee	Totalprice
New York	1	1	199
New York	1	2	2048
New York	3	2	18
Wuppertal	5	3	9

SalesTotal

```sql
CREATE VIEW SalesTotal AS
SELECT Location AS Shop, Month, SSN AS Employee,
       Price * Amount AS Totalprice
FROM Employee E, Shop H, Item I, Sales S
WHERE E.WorksFor = H.Location
      AND E.SSN = S.Employee
      AND I.Id = S.Item
```

Employee

SSN	Name	WorksFor
123	Peter Peterson	New York
342	Jane Janeson	New York
555	Heinz Heinzmann	Wuppertal

Shop

Location	Budget
New York	1.000.000
Wuppertal	4.000

Item

Id	Description	Price
1	Lawnmower	199
2	Fertilizer	32
3	Rake	9

Sales

Employee	Item	Amount	Month
123	1	1	1
342	2	64	1
342	3	2	3
555	3	1	5

Debugging

Origin of Result Tuples

- Tuple in result suspicious/wrong/interesting
- Learn more by looking at relevant inputs (provenance)

Example

SalesTotal

Shop	Month	Employee	Totalprice
New York	1	1	199
New York	1	2	2048
New York	3	2	18
Wuppertal	5	3	9

SalesTotal

```
CREATE VIEW SalesTotal AS
SELECT Location AS Shop, Month, SSN AS Employee,
       Price * Amount AS Totalprice
FROM Employee E, Shop H, Item I, Sales S
WHERE E.WorksFor = H.Location
      AND E.SSN = S.Employee
      AND I.Id = S.Item
```

Employee

SSN	Name	WorksFor
123	Peter Peterson	New York
342	Jane Janeson	New York
555	Heinz Heinzmann	Wuppertal

Shop

Location	Budget
New York	1.000.000
Wuppertal	4.000

Item

Id	Description	Price
1	Lawnmower	199
2	Fertilizer	32
3	Rake	9

Sales

Employee	Item	Amount	Month
123	1	1	1
342	2	64	1
342	3	2	3
555	3	1	5

Debugging

Approach

1. Identify tuples of interest (I)

ILLINOIS INSTITUTE OF TECHNOLOGY

Debugging

Approach

1. Identify tuples of interest (I)
2. Retrieve provenance
   - Need system that returns provenance for set I

Debugging

Approach

1. Identify tuples of interest (I)
2. Retrieve provenance
   - Need system that returns provenance for set I
   - How to represent this info?

ILLINOIS INSTITUTE OF TECHNOLOGY

Debugging

Approach

1. Identify tuples of interest (I)
2. Retrieve provenance
   - Need system that returns provenance for set I
   - How to represent this info?
3. What if provenance large?
   - $\Rightarrow$Query support? Visualization?

ILLINOIS INSTITUTE OF TECHNOLOGY

Annotation Propagation

Example

EnzymeProduce

Enzyme	Gene	
EC 1.1.1.1	ALB	??
EC 1.97.1.6	ALB	??

```
CREATE VIEW EnzymeProduce AS
SELECT Enzyme, Name AS Gene
FROM Gene G, Enzyme E
WHERE G.Id = E.ProducedBy
```

Gene

Id	Name	
4q11-q13	ALB	$\{a_4\}$
18q21.3	BCL2	$\{\}$

Enzyme

Enzyme	Weight	ProducedBy	
EC 1.1.1.1	45	4q11-q13	$\{a_1, a_2\}$
EC 1.97.1.6	12	4q11-q13	$\{a_2, a_3\}$

a_1 Necessary for red blood cells
a_2 Produced in liver
a_3 Unhealthy
a_4 Discovered by Edmond Hillary

Annotation Propagation

Which annotations in query result?

- Find provenance for tuple
- Attach union of annotations in provenance

Example

EnzymeProduce

Enzyme	Gene	
EC 1.1.1.1	ALB	??
EC 1.97.1.6	ALB	??

```
CREATE VIEW EnzymeProduce AS
SELECT Enzyme, Name AS Gene
FROM Gene G, Enzyme E
WHERE G.Id = E.ProducedBy
```

Gene

Id	Name	
4q11-q13	ALB	$\{a_4\}$
18q21.3	BCL2	$\{\}$

Enzyme

Enzyme	Weight	ProducedBy	
EC 1.1.1.1	45	4q11-q13	$\{a_1, a_2\}$
EC 1.97.1.6	12	4q11-q13	$\{a_2, a_3\}$

Annotation Propagation

For Example

- First result tuple
- **Provenance:** first tuples from `Gene` and `Enzyme`
- **Annotation:** a_1, a_2, a_4

Example

EnzymeProduce

Enzyme	Gene	
EC 1.1.1.1	ALB	??
EC 1.97.1.6	ALB	??

```sql
CREATE VIEW EnzymeProduce AS
SELECT Enzyme, Name AS Gene
FROM Gene G, Enzyme E
WHERE G.Id = E.ProducedBy
```

Gene

Id	Name	
4q11-q13	ALB	$\{a_4\}$
18q21.3	BCL2	$\{\}$

Enzyme

Enzyme	Weight	ProducedBy	
EC 1.1.1.1	45	4q11-q13	$\{a_1, a_2\}$
EC 1.97.1.6	12	4q11-q13	$\{a_2, a_3\}$

Annotation Propagation - Caveats

Potential Problems?

- What about negative influence?
- User should have control on propagation?
- What about annotations on
  - Attribute values
  - Spanning several tuples/relations/attributes

ILLINOIS INSTITUTE OF TECHNOLOGY

Deletion Propagation

Problem

- Given a *materialized view*

ILLINOIS INSTITUTE OF TECHNOLOGY

Deletion Propagation

Problem

- Given a *materialized view*
  - Query result stored as a table

Deletion Propagation

Problem

- Given a *materialized view*
  - Query result stored as a table
- How to update the view when input data changes

ILLINOIS INSTITUTE OF TECHNOLOGY

Deletion Propagation

Problem

- Given a *materialized view*
  - Query result stored as a table
- How to update the view when input data changes
  - Without recomputing the whole query

Deletion Propagation

Problem

- Given a *materialized view*
  - Query result stored as a table
- How to update the view when input data changes
  - Without recomputing the whole query
- **Deletion Propagation**: Update the view when input tuples are deleted?

Deletion Propagation Example

Example

```
CREATE VIEW ActiveCS AS
SELECT DISTINCT E.Name AS Emp
FROM Employee E, Project P, Assigned A
WHERE E.Id = A.Emp AND P.Name = A.Project
      AND Dep = CS
```

Employee

	Id	Name
e_1	1	Peter
e_2	2	Gertrud
e_2	3	Michael

Project

	Name	Dep
p_1	Server	CS
p_2	Webpage	CS
p_3	Fire CS	HR

Assigned

	Project	Emp
a_1	Server	1
a_2	Server	2
a_3	Webpage	2
a_4	Fire CS	3

Deletion Propagation Example

Example

ActiveCS

	Emp
t_1	Peter
t_2	Gertrud

Employee

	Id	Name
e_1	1	Peter
e_2	2	Gertrud
e_2	3	Michael

Project

	Name	Dep
p_1	Server	CS
p_2	Webpage	CS
p_3	Fire CS	HR

Assigned

	Project	Emp
a_1	Server	1
a_2	Server	2
a_3	Webpage	2
a_4	Fire CS	3

ILLINOIS INSTITUTE OF TECHNOLOGY

Deletion Propagation Example

Example

- Delete tuple from `Projects`

ActiveCS

	Emp
t_1	Peter
t_2	Gertrud

Employee

	Id	Name
e_1	1	Peter
e_2	2	Gertrud
e_2	3	Michael

Project

	Name	Dep
p_1	Server	CS
p_2	Webpage	CS
p_3	Fire CS	HR

Assigned

	Project	Emp
a_1	Server	1
a_2	Server	2
a_3	Webpage	2
a_4	Fire CS	3

Deletion Propagation Example

Example

- What would be the effect on the view?

ActiveCS

	Emp
t_1	Peter
t_2	Gertrud

Employee

	Id	Name
e_1	1	Peter
e_2	2	Gertrud
e_2	3	Michael

Project

	Name	Dep
p_1	Server	CS
p_2	Webpage	CS
p_3	Fire CS	HR

Assigned

	Project	Emp
a_1	Server	1
a_2	Server	2
a_3	Webpage	2
a_4	Fire CS	3

Deletion Propagation - Approach

Assumption

- Assume we have provenance for each tuple
  - For now a *set* of input tuples
  - $P(t_1) = \{e_1, p_1, a_1\}$
  - $P(t_2) = \{e_2, p_1, p_2, a_2, a_3\}$
- Set of deleted tuples ($D = \{p_1\}$)

ILLINOIS INSTITUTE OF TECHNOLOGY

Deletion Propagation - Approach

Assumption

- Assume we have provenance for each tuple
  - For now a *set* of input tuples
  - $P(t_1) = \{e_1, p_1, a_1\}$
  - $P(t_2) = \{e_2, p_1, p_2, a_2, a_3\}$
- Set of deleted tuples ($D = \{p_1\}$)

Approach

1. Remove D from provenance
2. Remove tuples without justification from view
   - Set provenance model to simple
   - Will learn later how this actually works

Deletion Propagation Example

Example

- $P(t_1) = \{e_1, p_1, a_1\} \rightarrow \{e_1, a_1\}$
- $P(t_2) = \{e_2, p_1, p_2, a_2, a_3\} \rightarrow \{e_2, p_2, a_2, a_3\}$

ActiveCS

	Emp
t_1	Peter
t_2	Gertrud

Employee

	Id	Name
e_1	1	Peter
e_2	2	Gertrud
e_2	3	Michael

Project

	Name	Dep
p_1	Server	CS
p_2	Webpage	CS
p_3	Fire CS	HR

Assigned

	Project	Emp
a_1	Server	1
a_2	Server	2
a_3	Webpage	2
a_4	Fire CS	3

Outline

Generation

Manual vs. Automatic

- **Manual**: User has to provide provenance information
- **Automatic**: System generates provenance information automatically
- Design space: How much information has to be provided by the user or transformation developer?

Lazy vs. Eager

- **Eager**: Generate provenance while the transformation is running
- **Lazy**: Generate provenance later once it is requested
- Tradeoff: Retrieval time vs. execution overhead

Generation

Approaches

- Run transformation in supervised environment that tracks provenance
- Instrument the transformations to produce provenance
- Record some information during execution and reconstruct provenance from this information

Supervised Environment

Idea

- Modify execution environment of transformations to capture provenance

Considerations

- What provenance to capture?
- Which parts of system ...
  - Are accessible?
  - Are modifiable?
- Supervision for all or only some transformations

ILLINOIS INSTITUTE OF TECHNOLOGY

Supervised Environment - Example

Example

- Hadoop - Map/Reduce
- Modify the Hadoop system to
  - store relationships between input/output keys
  - for mappers and reducers
  - in HDFS?

Supervised Environment - Discussion

Advantages

- Can capture whatever provenance we want
- No modification to transformations

Disadvantages

- **Intrusive** May have to re-implement whole system
- Overhead for transformation execution
- Parts of the system may not be accessible (e.g., web-service composition)

ILLINOIS INSTITUTE OF TECHNOLOGY

Instrument Transformations

Idea

- Modify the transformation to track its own provenance

Considerations

- Transformation language expressive enough to compute its own provenance?
- How to represent provenance in the data model?

ILLINOIS INSTITUTE OF TECHNOLOGY

Instrument Transformations - Example

Example

- SQL queries
- Rewrite queries to produce their output + provenance information
- Possible?

Instrument Transformations - Example

Example

- SQL queries
- Rewrite queries to produce their output + provenance information
- Possible? yes, later in course

Instrument Transformations - Example

Example

- SQL queries
- Rewrite queries to produce their output + provenance information
- Possible? yes, later in course
- Build a middleware that does that over standard DBMS

ILLINOIS INSTITUTE OF TECHNOLOGY

Instrument Transformations - Discussion

Advantages

- **Non-intrusive**: Possible without changes to system
  - If we can gather enough information about transformation from outside
  - E.g., DBMS client
- No overhead if no provenance computed
- Same data model ⇒Querying
- No manual changes to transformations

Disadvantages

- Performance optimizations may be limited (overhead provenance computation)
- Data model may limit the provenance representation

Reconstruction

Idea

- Recover provenance from input + output data and knowledge about transformation

Considerations

- Possible to know what's going on in the black box?
- Need to store extra information

ILLINOIS INSTITUTE OF TECHNOLOGY

Reconstruction - Example

Example

- Simple SQL query
- Write program to
  - Analyse query
  - Retrieve input and output data
  - Compute provenance

Reconstruction - Discussion

Advantages

- **Non-intrusive**: No changes to system
- No overhead for transformation
- No storage costs or almost no storage

Disadvantages

- Not possible for complex operations
- Provenance generation may be more expensive

ILLINOIS INSTITUTE OF TECHNOLOGY

Eager Generation

Approach

- Generate provenance during transformation execution

Considerations

- Overhead for transformation?
- How to trigger?

Lazy Generation

Approach

- Generate provenance on request

Considerations

- Input/Output data still available?
- Transformation info available?

ILLINOIS INSTITUTE OF TECHNOLOGY

Storage

- Provenance data can be orders of magnitude larger than input/output data
- ⇒ Be clever when to store what at which level of abstraction
- ⇒ Specialized compression for provenance
- ⇒ Index structured for provenance specific retrieval patterns

ILLINOIS INSTITUTE OF TECHNOLOGY

Why is provenance large?

Simplified explanation:

ILLINOIS INSTITUTE OF TECHNOLOGY

Why is provenance large?

Simplified explanation:

- Input data: size N

ILLINOIS INSTITUTE OF TECHNOLOGY

Why is provenance large?

Simplified explanation:

- Input data: size N
- Output data: size M

ILLINOIS INSTITUTE OF TECHNOLOGY

Why is provenance large?

Simplified explanation:

- Input data: size N
- Output data: size M
- Provenance is relationship between inputs and outputs

Why is provenance large?

Simplified explanation:

- Input data: size N
- Output data: size M
- Provenance is relationship between inputs and outputs
- $\Rightarrow$Worst case: $N \times M$

ILLINOIS INSTITUTE OF TECHNOLOGY

Why is provenance large?

Simplified explanation:

- Input data: size N
- Output data: size M
- Provenance is relationship between inputs and outputs
- $\Rightarrow$Worst case: $N \times M$
- Intermediate results?

ILLINOIS INSTITUTE OF TECHNOLOGY

Why is provenance large?

Simplified explanation:

- Input data: size N
- Output data: size M
- Provenance is relationship between inputs and outputs
- $\Rightarrow$Worst case: $N \times M$
- Intermediate results?
  - Transformation is tree with X nodes

ILLINOIS INSTITUTE OF TECHNOLOGY

Why is provenance large?

Simplified explanation:

- Input data: size N
- Output data: size M
- Provenance is relationship between inputs and outputs
- $\Rightarrow$Worst case: $N \times M$
- Intermediate results?
  - Transformation is tree with X nodes
  - $\Rightarrow \sim N \times M \times X$

ILLINOIS INSTITUTE OF TECHNOLOGY

What to store and when?

What?

- Only necessary level of detail
  - E.g., need attribute level provenance?
  - E.g., need provenance for intermediate results?

When?

- Provenance for all transformations?
- Only for specific type?
- Only when requested by user?
- Only when triggering event happend?

ILLINOIS INSTITUTE OF TECHNOLOGY

Compression

Rationale

- Provenance large, but has overlap
  - Exploit overlap to compress
  - Information loss?
  - Access/querying without decompression
  - Tradeoff: speed vs. size

Approaches

- Generic compression algorithms
  - Small size, slow?, probably no query
- Methods exploiting overlap being aware of provenance structure
  - Size less predictable, fast?, query may be possible

Index structures

Rationale

- Provenance querying needs efficient access to provenance data
- Traditional index structure useful?
- Can identify new access patterns?
  - Tree-path traversal?
- Static index or updates possible?

Approaches

- E.g., adapt IR retrieval index structures

ILLINOIS INSTITUTE OF TECHNOLOGY

Example Storage - Provenance tables

- Provenance Table
- input TID's → output TID's

Example

result

	shop	rev
t_1	Migros	125
t_2	Coop	25

↑

```sql
SELECT shop,
    sum(price) AS rev
FROM sales, items
WHERE itemId = id
GROUP BY shop
```

↑ ↑

sales

	shop	itemId
s_1	Migros	1
s_2	Migros	3
s_3	Coop	3

items

	id	price
i_1	1	100
i_2	2	10
i_3	3	25

Example Storage - Provenance tables

- Provenance Table
- input TID's $\rightarrow$ output TID's

Provenance

result	in
t_1	s_1
t_1	s_2
t_1	i_1
t_1	i_3
t_2	s_3
t_2	i_3

Example

result

	shop	rev
t_1	Migros	125
t_2	Coop	25

↑

```sql
SELECT shop,
    sum(price) AS rev
FROM sales, items
WHERE itemId = id
GROUP BY shop
```

↑ ↑

sales

	shop	itemId
s_1	Migros	1
s_2	Migros	3
s_3	Coop	3

items

	id	price
i_1	1	100
i_2	2	10
i_3	3	25

Querying

- Large amount of provenance information

Querying

- Large amount of provenance information
- Query support to extract information
  - Focus on parts of interest
    - Backward: Which data contributed to output?
    - Forward: Which data is derived from input?
    - Transitive closure
  - Correlated with input/output data
  - Summarize, abstract

ILLINOIS INSTITUTE OF TECHNOLOGY

Querying

- Large amount of provenance information
- Query support to extract information
  - Focus on parts of interest
    - Backward: Which data contributed to output?
    - Forward: Which data is derived from input?
    - Transitive closure
  - Correlated with input/output data
  - Summarize, abstract

Example

For a subset of erroneous sales totals, which ones have been derived from input sales data from a shop in New York with a amount sold bigger than 100.

Querying

Approaches

- Extend query language for "normal" data
- New query language

ILLINOIS INSTITUTE OF TECHNOLOGY

Querying

Approaches

- Extend query language for “normal” data
  - Querying provenance in combination with “normal” data
  - Limitation on provenance representation
- New query language

ILLINOIS INSTITUTE OF TECHNOLOGY

Querying

Approaches

- Extend query language for "normal" data
  - Querying provenance in combination with "normal" data
  - Limitation on provenance representation
- New query language
  - Operations tailored for typical operations on provenance
  - "Re-inventing the wheel"

ILLINOIS INSTITUTE OF TECHNOLOGY

Querying Example

Example

Provenance

result	in
t_1	s_1
t_1	s_2
t_1	i_1
t_1	i_3
t_2	s_3
t_2	i_3

Example

result

	shop	rev
t_1	Migros	125
t_2	Coop	25

↑

```
SELECT shop,
    sum(price) AS rev
FROM sales, items
WHERE itemId = id
GROUP BY shop
```

↑ ↑

sales

	shop	itemId
s_1	Migros	1
s_2	Migros	3
s_3	Coop	3

items

	id	price
i_1	1	100
i_2	2	10
i_3	3	25

Querying Example

```
SELECT DISTINCT shop
FROM result r,
     Provenance p,
     items i
WHERE r.tid = p.result
      AND p.in = i.tid
      AND i.price > 90
```

Example

Provenance

result	in
t_1	s_1
t_1	s_2
t_1	i_1
t_1	i_3
t_2	s_3
t_2	i_3

Example

result

	shop	rev
t_1	Migros	125
t_2	Coop	25

↑

```
SELECT shop,
    sum(price) AS rev
FROM sales, items
WHERE itemId = id
GROUP BY shop
```

↑ ↑

sales

	shop	itemId
s_1	Migros	1
s_2	Migros	3
s_3	Coop	3

items

	id	price
i_1	1	100
i_2	2	10
i_3	3	25

Querying Example

Prov query

	shop
q_1	Migros

Example

Provenance

result	in
t_1	s_1
t_1	s_2
t_1	i_1
t_1	i_3
t_2	s_3
t_2	i_3

Example

result

	shop	rev
t_1	Migros	125
t_2	Coop	25

↑

```sql
SELECT shop,
    sum(price) AS rev
FROM sales, items
WHERE itemId = id
GROUP BY shop
```

↑ ↑

sales

	shop	itemId
s_1	Migros	1
s_2	Migros	3
s_3	Coop	3

items

	id	price
i_1	1	100
i_2	2	10
i_3	3	25

Outline

ILLINOIS INSTITUTE OF TECHNOLOGY

Types of Provenance

- Data
- Transformation
- Other
- **Granularities**

Generation, Storage, and Querying

- **Generation**
  - Manual vs. automatic
  - Eager vs. lazy
  - Supervised environment, instrumentation, reconstruction
- **Storage**
  - Compression, Indices, What to keep?
- **Querying**
  - Extending transformation language
  - Develop new query language

Surveys I

Boris Glavic and Renée J. Miller.
Reexamining Some Holy Grails of Data Provenance.
In Tapp '11: 3rd usenix workshop on the theory and practice of provenance, 2011.

James Cheney, Laura Chiticariu, and Wang-Chiew Tan.
Provenance in Databases: Why, How, and Where.
Foundations and Trends in Databases, 1(4):379–474, 2009.

Robert Ikeda and Jennifer Widom.
Data Lineage: A Survey.
Technical report, Stanford University, 2009.

Juliana Freire, David Koop, Emanuele Santos, and Claudio T. Silva.
Provenance for Computational Tasks: A Survey.
Computing in Science and Engineering, 10(3):11–21, 2008.

Boris Glavic and Klaus R. Dittrich.
Data Provenance: A Categorization of Existing Approaches.
In Btw '07: proceedings of datenbanksysteme in buisness, technologie und web, 227-241, 2007.

Susan B. Davidson, Sarah Cohen-Boulakia, Anat Eyal, Bertram Ludscher, Timothy McPhillips, Shawn Bowers, and Juliana Freire.
Provenance in Scientific Workflow Systems.
IEEE Data Engineering Bulletin, 32(4):44–50, 2007.

ILLINOIS INSTITUTE OF TECHNOLOGY

Surveys II

Wang-Chiew Tan.
Provenance in Databases: Past, Current, and Future.
IEEE Data Engineering Bulletin, 30(4):3–12, 2007.

Yogesh L. Simmhan, Beth Plale, and Dennis Gannon.
A Survey of Data Provenance in e-science.
SIGMOD Record, 34(3):31–36, 2005.

Yogesh L. Simmhan, Beth Plale, and Dennis Gannon.
A Survey of Data Provenance Techniques.
Technical report, Indiana University, Bloomington IN 47405, 2005.

Wang-Chiew Tan.
Research Problems in Data Provenance.
IEEE Data Engineering Bulletin, 27(4):42-52, 2004.

ILLINOIS INSTITUTE OF TECHNOLOGY