CSE 444

Lecture 28:

Provenance

Announcements

- Quiz section on Thursday: canceled
- Lecture on Friday: canceled
- Lab 4 / Lab 6: due on Friday night
- Final writeup: on on Saturday night

- UW Course Evaluations:
  – Online https://uw.iasystem.org/survey/130405
  – Until June 12, 2014

Data Provenance

Data Provenance

- Provenance inside the DBMS
  - Will discuss today

- Provenance outside of the DBMS
  - In workflows: keep track of which dataset was produced by what program, which version, on what date, and using what input data
  - Much more messy; there is a standard, OPM (Open Provenance Model)

Provenance Annotations

- Some query produces an output table T(A,B,C)
- We store it over some period of time
- Later we ask: “where did this tuple come from?”
- The “provenance annotation” answers this.

A	B	C	
a1	b1	c1	provenance1
a2	b1	c1	provenance2
a2	b2	c2	provenance3
a2	b2	c3	provenance4

Provenance Annotations

- Start by annotating each tuple in the original database with a unique identifier; can be the Tuple Id (TID)

A	B	
a1	b1	X1
a2	b1	X2
a2	b2	X3

- Next, compute the provenance expression inductively, based on the query plan

Join Operator

$\bowtie$

A	B	
a1	b1	X1
a2	b1	X2
a2	b2	X3

B	C	
b1	c1	Y1
b2	c2	Y2
b2	c3	Y3

=

A	B	C	
a1	b1	c1	X1 · Y1
a2	b1	c1	X2 · Y1
a2	b2	c2	X3 · Y2
a2	b2	c3	X3 · Y3

Projection Operator

Π

A	B	
a1	b1	X1
a1	b2	X2
a2	b1	X3
a2	b2	X4
a2	b3	X5

=

A	
a1	X1+X2
a2	X3+X4+X5

Union Operator

$\cup$

A	B	
a1	b1	X1
a2	b2	X2

A	B	
a2	b2	Y1
a3	b3	Y2

=

A	B	
a1	b1	X1
a2	b2	X2+Y1
a3	b3	X3

Selection Operator

$\sigma_{A=a1}$

A	B	
a1	b1	X1
a1	b2	X2
a2	b1	X3
a2	b2	X4
a2	b3	X5

=

A	B	
a1	b1	X1
a1	b2	X2

We could simply remove the tuples filtered out.
But it's better to keep them around (we'll see why).
What is their annotation?

Selection Operator

$\sigma_{A=a1}$

A	B	
a1	b1	X1
a1	b2	X2
a2	b1	X3
a2	b2	X4
a2	b3	X5

=

A	B	
a1	b1	$X1 \cdot 1$
a1	b2	$X2 \cdot 1$
a2	b1	$X3 \cdot 0$
a2	b2	$X4 \cdot 0$
a2	b3	$X5 \cdot 0$

We could simply remove the tuples filtered out.
But it's better to keep them around (we'll see why).
What is their annotation?

Simple Example 1

$$\Pi_{AC}(R) \bowtie \Pi_{BC}(R) =$$

R

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

Simple Example 1

$$\Pi_{AC}(R) \bowtie \Pi_{BC}(R) =$$

R

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

$\Pi_{AC}(R)$

A	C	
a	c	X
d	e	Y
f	e	Z

Simple Example 1

$\Pi_{AC}(R) \bowtie \Pi_{BC}(R) =$

R

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

$\Pi_{AC}(R)$

A	C	
a	c	X
d	e	Y
f	e	Z

$\Pi_{BC}(R)$

B	C	
b	c	X
b	e	Y
g	e	Z

Simple Example 1

$$\Pi_{AC}(R) \bowtie \Pi_{BC}(R) =$$

R

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

$\Pi_{AC}(R)$

A	C	
a	c	X
d	e	Y
f	e	Z

$\Pi_{BC}(R)$

B	C	
b	c	X
b	e	Y
g	e	Z

A	B	C	
a	b	c	X · X
d	b	e	Y · Y
d	g	e	Y · Z
f	b	e	Z · Y
f	g	e	Z · Z

Simple Example 2

$\sigma_{C=e}(R) =$

R

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

A	B	C	
a	b	c	$0 = X \cdot 0$
d	b	e	$Y = Y \cdot 1$
f	g	e	$Z = Z \cdot 1$

Complex Example

$$\sigma_{C=e}\Pi_{AC}(\ \Pi_{AC}(R) \bowtie \Pi_{BC}(R) \cup \Pi_{AB}(R) \bowtie \Pi_{BC}(R)) =$$

R

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

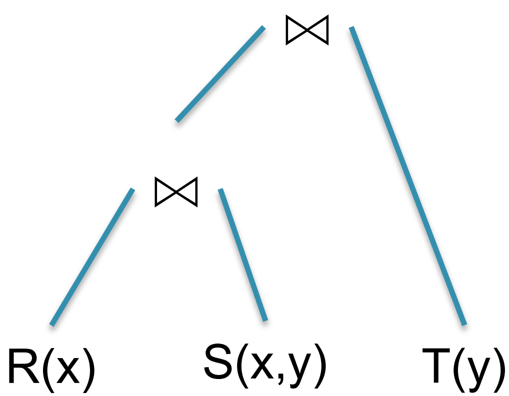
A	C	
a	c	$(X \cdot X + X \cdot X) \cdot 0 = 0 \cdot 2 \cdot X^2 = 0$
a	e	$X \cdot Y \cdot 1 = X \cdot Y$
d	c	$Y \cdot X \cdot 0 = 0$
d	e	$(Y \cdot Y + Y \cdot Z + Y \cdot Y) \cdot 1 = 2 \cdot Y^2 + Y \cdot Z$
f	e	$(Z \cdot Z + Z \cdot Y + Z \cdot Z) \cdot 1 = 2 \cdot Z^2 + Y \cdot Z$

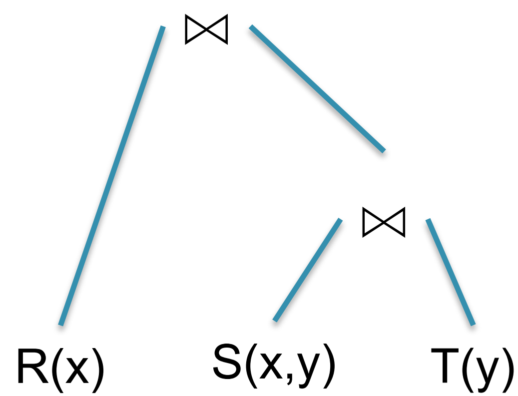
Discuss in class what these annotations mean

Independence of Plan

$q(x,y) := R(x), S(x,y), T(y)$

Do these plans compute the same provenance for the output (a,b)?





R=

x	
a	X

S=

x	y	
a	b	Y

T=

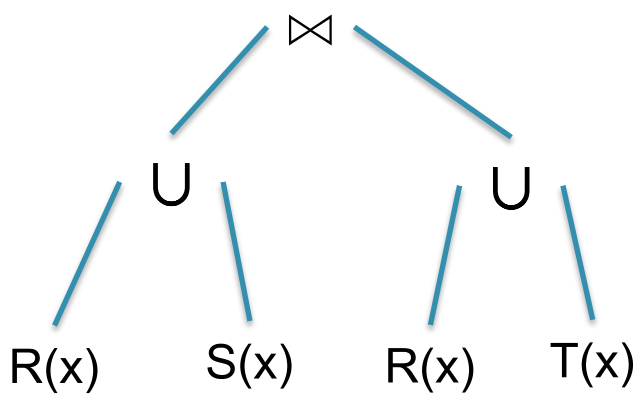
y	
b	Z

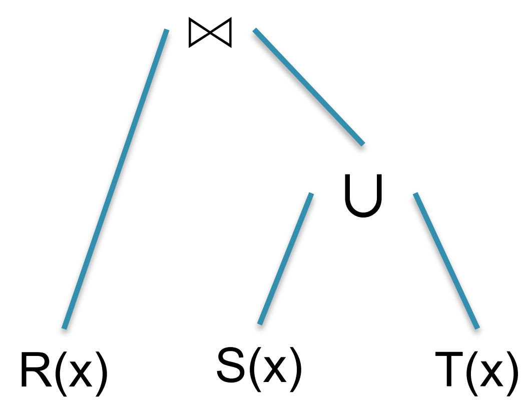
Independence of Plan

q(x) := R(x), S(x)
q(x) := R(x), T(x)

Do these two plans compute the same provenance expression for the output (a)?

V(x) := S(x)
V(x) := T(x)
q(x) := R(x), V(x)





R=

x
a

X

S=

x
a

Y

T=

x
a

Z

Identities on Provenance Expressions

Definition. A structure $(K, +, \cdot, 0, 1)$ is called a commutative semiring if:

1. $(K,+,0)$ is a commutative monoid:
   a. $+$ is associative: $(x+y)+z=x+(y+z)$
   b. $+$ is commutative: $x+y=y+x$
   c. 0 is the identity for $+$: $x+0=0+x=x$
2. $(K, \cdot, 1)$ is a commutative monoid:
   a. … (similar identities)
3. $\cdot$ distributes over $+$: $x \cdot (y+z) = x \cdot y + x \cdot z$
4. For all x: $x \cdot 0 = 0 \cdot x = 0$

Identities on Provenance Expressions

Definition. A structure $(K, +, \cdot, 0, 1)$ is called a commutative semiring if:

1. $(K,+,0)$ is a commutative monoid:
   a. $+$ is associative: $(x+y)+z=x+(y+z)$
   b. $+$ is commutative: $x+y=y+x$
   c. 0 is the identity for $+$: $x+0=0+x=x$
2. $(K, \cdot, 1)$ is a commutative monoid:
   a. … (similar identities)
3. $\cdot$ distributes over $+$: $x \cdot (y+z) = x \cdot y + x \cdot z$
4. For all x: $x \cdot 0 = 0 \cdot x = 0$

Fact: if we compute annotations in a commutative semiring, then the final result is the same for all plans that are equivalent under set semantics

Example

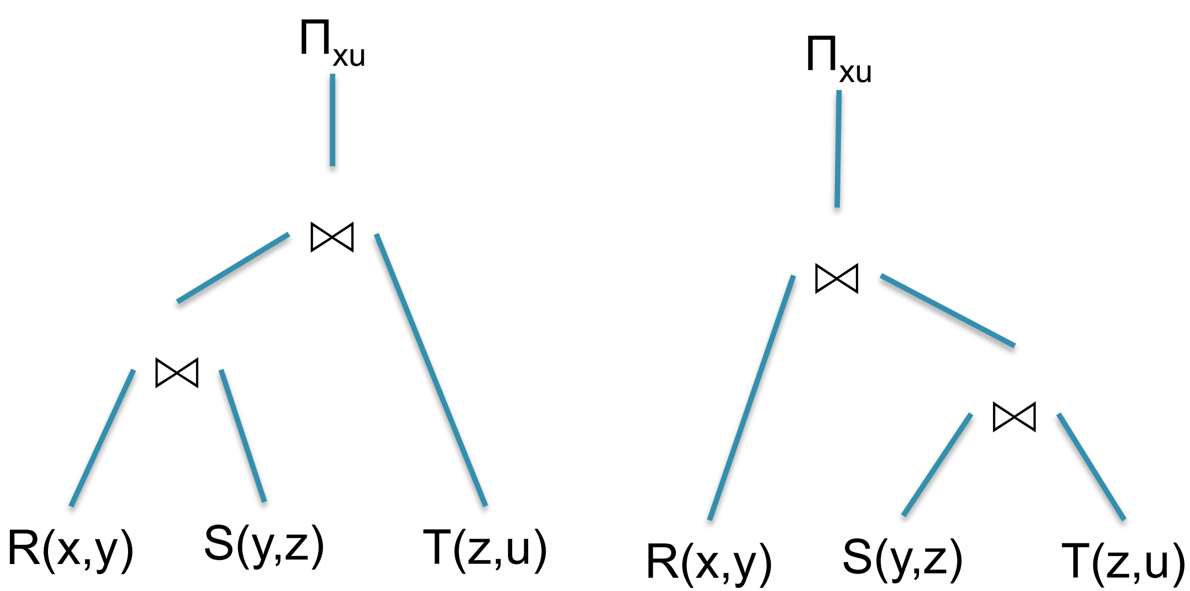
$q(x,u) := R(x,y), S(y,z), T(z,u)$

In class: compute the provenance of the output (a,b) for both plans.

x	y	
a	b1	X1
a	b2	X2

y	z	
b1	c1	Y1
b1	c2	Y2
b2	c2	Y3

z	u	
c1	d	Z1
c2	d	Z2



Applications

$$\sigma_{C=e}\Pi_{AC}(\ \Pi_{AC}(R) \bowtie \Pi_{BC}(R) \cup \Pi_{AB}(R) \bowtie \Pi_{BC}(R)) =$$

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

A	C	
a	c	0
a	e	$X \cdot Y$
d	e	$2 \cdot Y^2 + Y \cdot Z$
f	e	$2 \cdot Z^2 + Y \cdot Z$

Q: Suppose we delete the tuple (d,b,e) from R.
Which tuple(s) disappear from the result?

Applications

$$\sigma_{C=e}\Pi_{AC}(\ \Pi_{AC}(R) \bowtie \Pi_{BC}(R) \cup \Pi_{AB}(R) \bowtie \Pi_{BC}(R)) =$$

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

A	C	
a	c	0
a	e	$X \cdot Y$
d	e	$2 \cdot Y^2 + Y \cdot Z$
f	e	$2 \cdot Z^2 + Y \cdot Z$

=

A	C	
a	c	0
a	e	0
d	e	0
f	e	$2 \cdot Z^2$

Q: Suppose we delete the tuple (d,b,e) from R.
Which tuple(s) disappear from the result?

A: Set Y=0

Applications

$$\sigma_{C=e}\Pi_{AC}(\ \Pi_{AC}(R) \bowtie \Pi_{BC}(R) \cup \Pi_{AB}(R) \bowtie \Pi_{BC}(R)) =$$

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

A	C	
a	c	0
a	e	$X \cdot Y$
d	e	$2 \cdot Y^2 + Y \cdot Z$
f	e	$2 \cdot Z^2 + Y \cdot Z$

Q: Suppose each tuple in R occurs 3 times (bag semantics).
How many times occurs each tuple in the answer?

Applications

$$\sigma_{C=e}\Pi_{AC}(\ \Pi_{AC}(R) \bowtie \Pi_{BC}(R) \cup \Pi_{AB}(R) \bowtie \Pi_{BC}(R)) =$$

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

A	C	
a	c	0
a	e	$X \cdot Y$
d	e	$2 \cdot Y^2 + Y \cdot Z$
f	e	$2 \cdot Z^2 + Y \cdot Z$

A	C	
a	c	0
a	e	9
d	e	27
f	e	27

Q: Suppose each tuple in R occurs 3 times (bag semantics).
How many times occurs each tuple in the answer?

A. Set X=Y=Z=3

Application: A Simpler Provenance of Sets of Contributing Tuples

$$\sigma_{C=e}\Pi_{AC}(\,\Pi_{AC}(R) \bowtie \Pi_{BC}(R) \cup \Pi_{AB}(R) \bowtie \Pi_{BC}(R)) =$$

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

A	C	
a	c	0
a	e	$X \cdot Y$
d	e	$2 \cdot Y^2 + Y \cdot Z$
f	e	$2 \cdot Z^2 + Y \cdot Z$

➔

A	C	
a	c	-
a	e	X, Y
d	e	Y, Z
f	e	Y, Z

Trace only the set of input tuples that contributed to an output tuple

This is also a semi-ring! Which one?

Application: Security

Discretionary Access Control [LaPadula]

- Public = P
- Confidential = C
- Secret = S
- Top Secret = T
- No Such Thing… = 0

R =

A	B	C
a	b	c
d	b	e
f	g	e

X=C

Y=P

Z=T

A	C	
a	c	$2 \cdot X^2 =$?
a	e	$X \cdot Y =$?
d	e	$2 \cdot Y^2 + Y \cdot Z =$?
f	e	$2 \cdot Z^2 + Y \cdot Z =$?

Application: Security

Discretionary Access Control [LaPadula]
- Public = P
- Confidential = C
- Secret = S
- Top Secret = T
- No Such Thing… = 0

Alice has clearance S:
- Alice can read C data
- Alice cannot read T data

- Alice can write T data
- Alice cannot read C data

Why??

Application: Security

Discretionary Access Control [LaPadula]
- Public = P
- Confidential = C
- Secret = S
- Top Secret = T
- No Such Thing… = 0

Alice has clearance S:
- Alice can read C data
- Alice cannot read T data

- Alice can write T data
- Alice cannot read C data

Why??

Q: Join record A labeled C with record B labeled S. What is the label of (A,B)?

Application: Security

Discretionary Access Control [LaPadula]
- Public = P
- Confidential = C
- Secret = S
- Top Secret = T
- No Such Thing… = 0

Alice has clearance S:
- Alice can read C data
- Alice cannot read T data

- Alice can write T data
- Alice cannot read C data

Why??

Q: Join record A labeled C with record B labeled S. What is the label of (A,B)?
A: S

Application: Security

Discretionary Access Control [LaPadula]
- Public = P
- Confidential = C
- Secret = S
- Top Secret = T
- No Such Thing… = 0

Alice has clearance S:
- Alice can read C data
- Alice cannot read T data

- Alice can write T data
- Alice cannot read C data

Why??

Q: Join record A labeled C with record B labeled S. What is the label of (A,B)?
A: S

Q: Eliminate duplicates {A, A, A,A} labeled T, C, C, S. What is the label of A?

Application: Security

Discretionary Access Control [LaPadula]
- Public = P
- Confidential = C
- Secret = S
- Top Secret = T
- No Such Thing… = 0

Alice has clearance S:
- Alice can read C data
- Alice cannot read T data

- Alice can write T data
- Alice cannot read C data

Why??

Q: Join record A labeled C with record B labeled S. What is the label of (A,B)?
A: S

Q: Eliminate duplicates {A, A, A,A} labeled T, C, C, S. What is the label of A?
A: C

Application: Security

Discretionary Access Control [LaPadula]

- Public = P
- Confidential = C
- Secret = S
- Top Secret = T
- No Such Thing… = 0

What are the labels of these records?

R =

A	B	C
a	b	c
d	b	e
f	g	e

X=C

Y=P

Z=T

A	C	
a	c	$2 \cdot X^2$
a	e	$X \cdot Y$
d	e	$2 \cdot Y^2 + Y \cdot Z$
f	e	$2 \cdot Z^2 + Y \cdot Z$

(A, min, max, 0, P), where A = P < C < S < T < 0

Application: Security

Discretionary Access Control [LaPadula]

- Public = P
- Confidential = C
- Secret = S
- Top Secret = T
- No Such Thing… = 0

R =

A	B	C	
a	b	c	X=C
d	b	e	Y=P
f	g	e	Z=T

A	C	
a	c	$2 \cdot X^2 = C$
a	e	$X \cdot Y = C$
d	e	$2 \cdot Y^2 + Y \cdot Z = C$
f	e	$2 \cdot Z^2 + Y \cdot Z = T$

(A, min, max, 0, P), where A = P < C < S < T < 0

Summary

- In many applications it is critical to record the provenance of the data
- Fine grained provenance:
  - Inside the database system
  - Clear semantics that aligns to relational queries
  - This is what we discussed today
- Coarse grained provenance:
  - Lossy, by necessity
  - Trade off accuracy for size