Introduction to Data Management CSE 344

Lecture 18: Design Theory Wrap-up and Views

Announcements

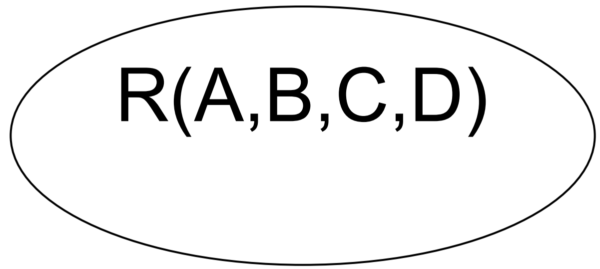
- Midterm is graded: mean ≈ median ≈ 63
  - Please read the solutions carefully.
- WQ7 is due tomorrow
- Homework 5 is due on Friday
- No classes on Monday – Veteran's day

- Today:
  - Finish design theory (FD, BCNF)
  - Views

R(A,B,C,D)

A → B
B → C

Review: BCNF



R(A,B,C,D)

$A \rightarrow B$
$B \rightarrow C$

Review: BCNF

Recall: find X s.t.
$X \subsetneq X^+ \subsetneq$ [all-attrs]

R(A,B,C,D)

R(A,B,C,D)

$A \rightarrow B$
$B \rightarrow C$

Review: BCNF

Recall: find X s.t.
$X \subsetneq X^+ \subsetneq$ [all-attrs]

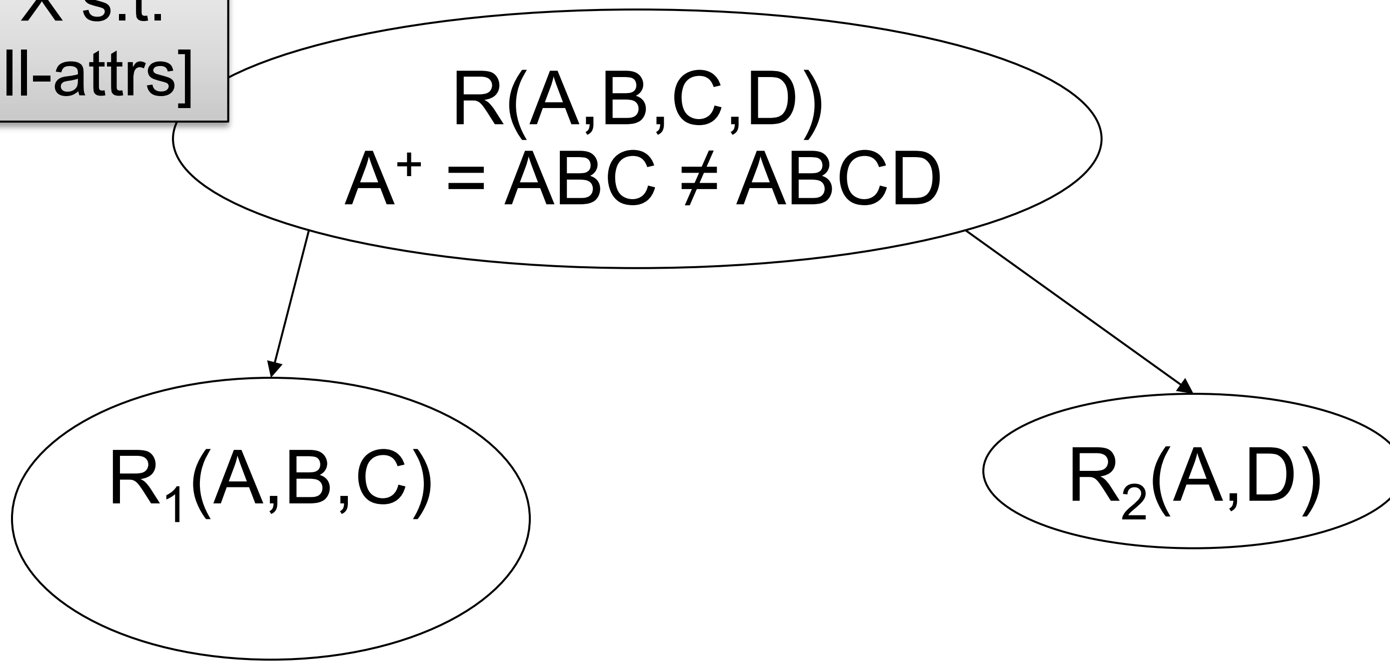
R(A,B,C,D)
$A^+ = ABC \neq ABCD$

R(A,B,C,D)

$A \rightarrow B$
$B \rightarrow C$

Review: BCNF

Recall: find X s.t.
$X \subsetneq X^+ \subsetneq$ [all-attrs]

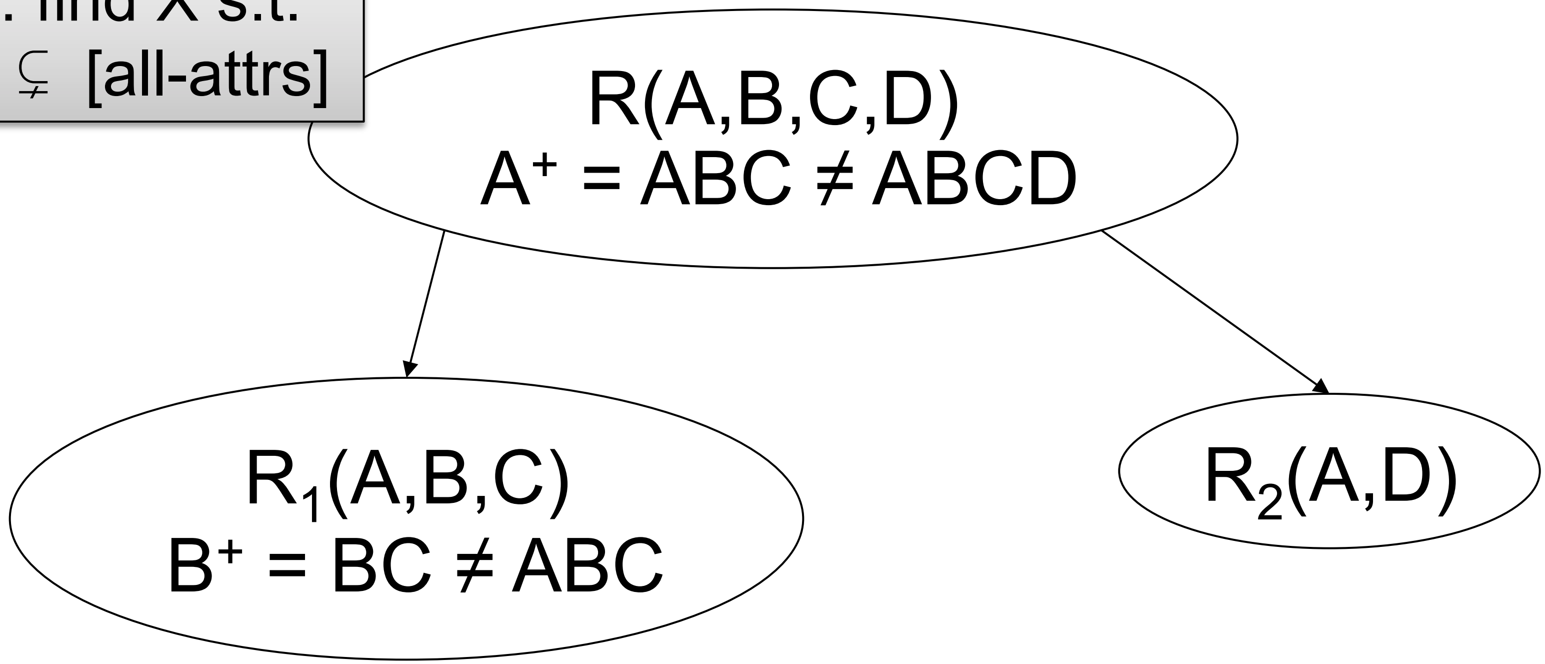


R(A,B,C,D)

A → B
B → C

Review: BCNF

Recall: find X s.t.
$X \subsetneq X^+ \subsetneq$ [all-attrs]

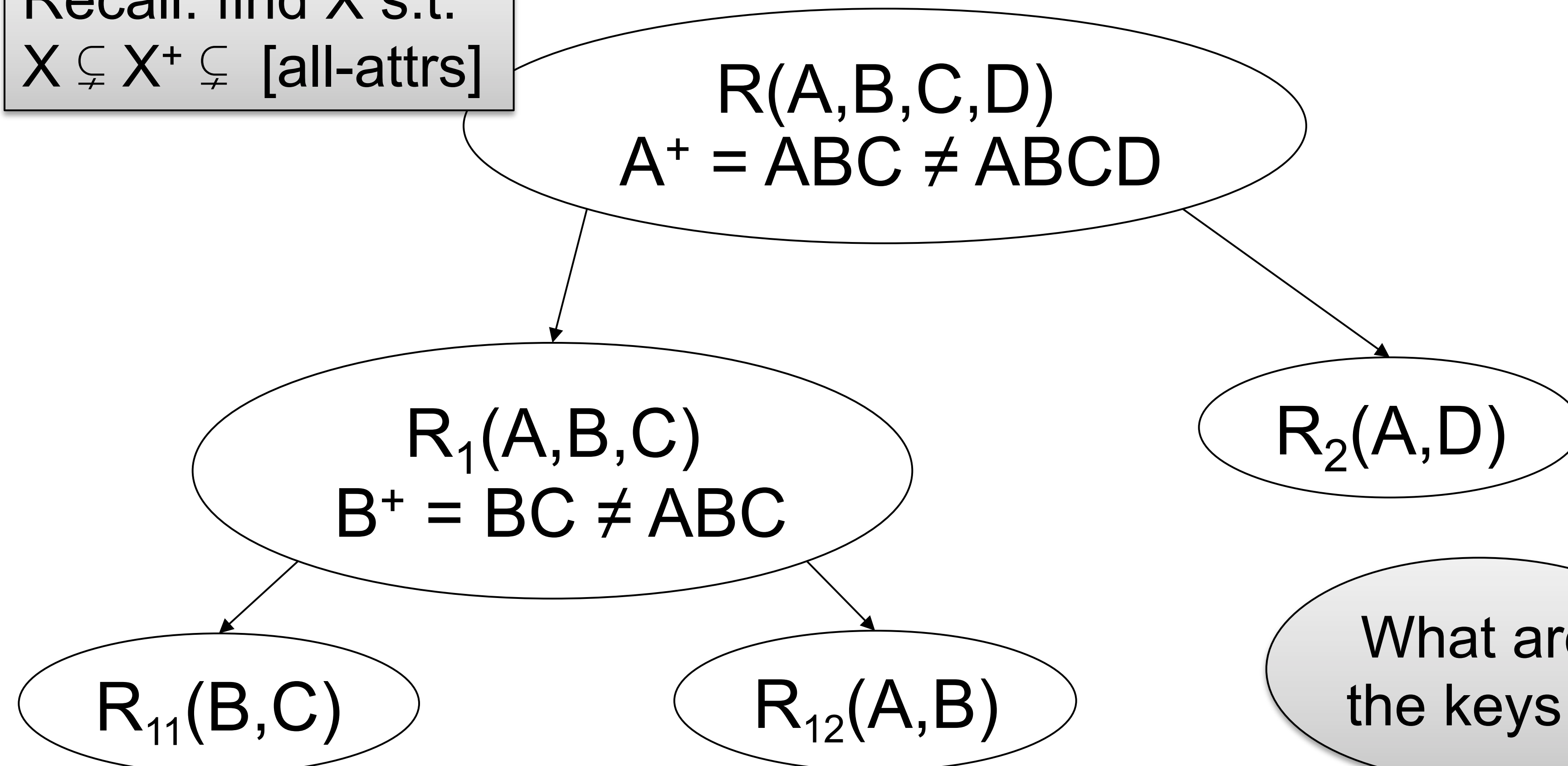


R(A,B,C,D)

A → B
B → C

Review: BCNF

Recall: find X s.t.
$X \subsetneq X^+ \subsetneq$ [all-attrs]



What are the keys ?

What happens if in R we first pick B^+ ? Or AB^+ ?

Decompositions in General

$R(A_1, \ldots, A_n, B_1, \ldots, B_m, C_1, \ldots, C_p)$

$S_1(A_1, \ldots, A_n, B_1, \ldots, B_m)$

$S_2(A_1, \ldots, A_n, C_1, \ldots, C_p)$

S_1 = projection of R on $A_1, \ldots, A_n, B_1, \ldots, B_m$
S_2 = projection of R on $A_1, \ldots, A_n, C_1, \ldots, C_p$

Lossless Decomposition

Name	Price	Category
Gizmo	19.99	Gadget
OneClick	24.99	Camera
Gizmo	19.99	Camera

Name	Price
Gizmo	19.99
OneClick	24.99
~~Gizmo~~	~~19.99~~

Name	Category
Gizmo	Gadget
OneClick	Camera
Gizmo	Camera

Lossy Decomposition

What is lossy here?

Name	Price	Category
Gizmo	19.99	Gadget
OneClick	24.99	Camera
Gizmo	19.99	Camera

Name	Category
Gizmo	Gadget
OneClick	Camera
Gizmo	Camera

Price	Category
19.99	Gadget
24.99	Camera
19.99	Camera

Decomposition in General

$R(A_1, ..., A_n, B_1, ..., B_m, C_1, ..., C_p)$

$S_1(A_1, ..., A_n, B_1, ..., B_m)$

$S_2(A_1, ..., A_n, C_1, ..., C_p)$

Let: S_1 = projection of R on $A_1, ..., A_n, B_1, ..., B_m$
S_2 = projection of R on $A_1, ..., A_n, C_1, ..., C_p$

The decomposition is called *lossless* if $R = S_1 \bowtie S_2$

Fact: If $A_1, ..., A_n \rightarrow B_1, ..., B_m$ then the decomposition is lossless

It follows that every BCNF decomposition is losselss

Example from textbook Ch. 3.4.2

The Chase Test for Lossless Join

$R(A,B,C,D) = S1(A,D) \bowtie S2(A,C) \bowtie S3(B,C,D)$
R satisfies: $A \rightarrow B$, $B \rightarrow C$, $CD \rightarrow A$

$S1 = \Pi_{AD}(R)$, $S2 = \Pi_{AC}(R)$, $S3 = \Pi_{BCD}(R)$,
hence $R \subseteq S1 \bowtie S2 \bowtie S3$
Need to check: $R \supseteq S1 \bowtie S2 \bowtie S3$

Example from textbook Ch. 3.4.2

The Chase Test for Lossless Join

$R(A,B,C,D) = S1(A,D) \bowtie S2(A,C) \bowtie S3(B,C,D)$
R satisfies: $A \rightarrow B$, $B \rightarrow C$, $CD \rightarrow A$

$S1 = \Pi_{AD}(R)$, $S2 = \Pi_{AC}(R)$, $S3 = \Pi_{BCD}(R)$,
hence $R \subseteq S1 \bowtie S2 \bowtie S3$
Need to check: $R \supseteq S1 \bowtie S2 \bowtie S3$
Suppose $(a,b,c,d) \in S1 \bowtie S2 \bowtie S3$ Is it also in R?
R must contain the following tuples:

A	B	C	D
a	b1	c1	d

Why ?
$(a,d) \in S1 = \Pi_{AD}(R)$

Example from textbook Ch. 3.4.2

The Chase Test for Lossless Join

R(A,B,C,D) = S1(A,D) ⋈ S2(A,C) ⋈ S3(B,C,D)
R satisfies: A→B, B→C, CD→A

S1 = Π_{AD}(R), S2 = Π_{AC}(R), S3 = Π_{BCD}(R),
hence R ⊆ S1 ⋈ S2 ⋈ S3

Need to check: R ⊇ S1 ⋈ S2 ⋈ S3

Suppose (a,b,c,d) ∈ S1 ⋈ S2 ⋈ S3 Is it also in R?

R must contain the following tuples:

A	B	C	D	Why ?
a	b1	c1	d	(a,d) ∈ S1 = Π_{AD}(R)
a	b2	c	d2	(a,c) ∈ S2 = Π_{BD}(R)

Example from textbook Ch. 3.4.2

The Chase Test for Lossless Join

$R(A,B,C,D) = S1(A,D) \bowtie S2(A,C) \bowtie S3(B,C,D)$
R satisfies: $A \rightarrow B$, $B \rightarrow C$, $CD \rightarrow A$

$S1 = \Pi_{AD}(R)$, $S2 = \Pi_{AC}(R)$, $S3 = \Pi_{BCD}(R)$,
hence $R \subseteq S1 \bowtie S2 \bowtie S3$
Need to check: $R \supseteq S1 \bowtie S2 \bowtie S3$
Suppose $(a,b,c,d) \in S1 \bowtie S2 \bowtie S3$ Is it also in R?
R must contain the following tuples:

A	B	C	D	Why ?
a	b1	c1	d	$(a,d) \in S1 = \Pi_{AD}(R)$
a	b2	c	d2	$(a,c) \in S2 = \Pi_{BD}(R)$
a3	b	c	d	$(b,c,d) \in S3 = \Pi_{BCD}(R)$

Example from textbook Ch. 3.4.2

The Chase Test for Lossless Join

R(A,B,C,D) = S1(A,D) ⋈ S2(A,C) ⋈ S3(B,C,D)
R satisfies: A→B, B→C, CD→A

S1 = Π_{AD}(R), S2 = Π_{AC}(R), S3 = Π_{BCD}(R),
hence R ⊆ S1 ⋈ S2 ⋈ S3
Need to check: R ⊇ S1 ⋈ S2 ⋈ S3
Suppose (a,b,c,d) ∈ S1 ⋈ S2 ⋈ S3 Is it also in R?
R must contain the following tuples:

A	B	C	D	Why ?
a	b1	c1	d	(a,d) ∈ S1 = Π_{AD}(R)
a	b2	c	d2	(a,c) ∈ S2 = Π_{BD}(R)
a3	b	c	d	(b,c,d) ∈ S3 = Π_{BCD}(R)

"Chase" them (apply FDs):

A→B

A	B	C	D
a	b1	c1	d
a	b1	c	d2
a3	b	c	d

Example from textbook Ch. 3.4.2

The Chase Test for Lossless Join

R(A,B,C,D) = S1(A,D) ⋈ S2(A,C) ⋈ S3(B,C,D)
R satisfies: A→B, B→C, CD→A

S1 = Π_{AD}(R), S2 = Π_{AC}(R), S3 = Π_{BCD}(R),
hence R ⊆ S1 ⋈ S2 ⋈ S3

Need to check: R ⊇ S1 ⋈ S2 ⋈ S3

Suppose (a,b,c,d) ∈ S1 ⋈ S2 ⋈ S3 Is it also in R?

R must contain the following tuples:

A	B	C	D	Why ?
a	b1	c1	d	(a,d) ∈ S1 = Π_{AD}(R)
a	b2	c	d2	(a,c) ∈ S2 = Π_{BD}(R)
a3	b	c	d	(b,c,d) ∈ S3 = Π_{BCD}(R)

"Chase" them (apply FDs):

A→B

A	B	C	D
a	b1	c1	d
a	b1	c	d2
a3	b	c	d

B→C

A	B	C	D
a	b1	c	d
a	b1	c	d2
a3	b	c	d

Example from textbook Ch. 3.4.2

The Chase Test for Lossless Join

R(A,B,C,D) = S1(A,D) ⋈ S2(A,C) ⋈ S3(B,C,D)
R satisfies: A→B, B→C, CD→A

S1 = Π_{AD}(R), S2 = Π_{AC}(R), S3 = Π_{BCD}(R),
hence R ⊆ S1 ⋈ S2 ⋈ S3

Need to check: R ⊇ S1 ⋈ S2 ⋈ S3

Suppose (a,b,c,d) ∈ S1 ⋈ S2 ⋈ S3 Is it also in R?

R must contain the following tuples:

A	B	C	D	Why ?
a	b1	c1	d	(a,d) ∈S1 = Π_{AD}(R)
a	b2	c	d2	(a,c) ∈S2 = Π_{BD}(R)
a3	b	c	d	(b,c,d) ∈S3 = Π_{BCD}(R)

"Chase" them (apply FDs):

A→B

A	B	C	D
a	b1	c1	d
a	b1	c	d2
a3	b	c	d

B→C

A	B	C	D
a	b1	c	d
a	b1	c	d2
a3	b	c	d

CD→A

A	B	C	D
a	b1	c	d
a	b1	c	d2
a	b	c	d

Hence R contains (a,b,c,d)

Schema Refinements = Normal Forms

- 1st Normal Form = all tables are flat
- 2nd Normal Form = obsolete
- Boyce Codd Normal Form = discussed in class
- 3rd Normal Form = see book

Views

- A view in SQL =
  - A table computed from other tables, s.t., whenever the base tables are updated, the view is updated too
- More generally:
  - A view is derived data that keeps track of changes in the original data
- Compare:
  - A function computes a value from other values, but does not keep track of changes to the inputs

Purchase(customer, product, store)
Product(pname, price)

StorePrice(store, price)

A Simple View

Create a view that returns for each store
the prices of products purchased at that store

```
CREATE VIEW  StorePrice AS
  SELECT DISTINCT x.store, y.price
  FROM  Purchase x, Product y
  WHERE x.product = y.pname
```

This is like a new table
StorePrice(store,price)

Purchase(customer, product, store)
Product(pname, price)

StorePrice(store, price)

We Use a View Like Any Table

- A "high end" store is a store that sell some products over 1000.
- For each customer, return all the high end stores that they visit.

```
SELECT DISTINCT u.name, u.store
FROM Purchase u, StorePrice v
WHERE u.store = v.store
    AND v.price > 1000
```

Types of Views

- Virtual views
  - Used in databases
  - Computed only on-demand – slow at runtime
  - Always up to date

- Materialized views
  - Used in data warehouses
  - Pre-computed offline – fast at runtime
  - May have stale data (must recompute or update)
  - Indexes *are* materialized views

Purchase(customer, product, store)
Product(pname, price)

StorePrice(store, price)

Query Modification

For each customer, find all the high end stores that they visit.

```
CREATE VIEW  StorePrice AS
  SELECT DISTINCT x.store, y.price
  FROM  Purchase x, Product y
  WHERE x.product = y.pname
```

```
SELECT DISTINCT u.name, u.store
FROM Purchase u, StorePrice v
WHERE u.store = v.store
   AND v.price > 1000
```

Purchase(customer, product, store)
Product(pname, price)

StorePrice(store, price)

Query Modification

For each customer, find all the high end stores that they visit.

```
CREATE VIEW  StorePrice AS
  SELECT DISTINCT x.store, y.price
  FROM  Purchase x, Product y
  WHERE x.product = y.pname
```

```
SELECT DISTINCT u.name, u.store
FROM Purchase u, StorePrice v
WHERE u.store = v.store
   AND v.price > 1000
```

Modified query:

```
SELECT DISTINCT u.customer, u.store
FROM Purchase u,
 (SELECT DISTINCT x.store, y.price
  FROM  Purchase x, Product y
  WHERE x.product = y.pname) v
WHERE u.store = v.store
    AND v.price > 1000
```

Purchase(customer, product, store)
Product(pname, price)

StorePrice(store, price)

Query Modification

For each customer, find all the high end stores that they visit.

```
SELECT DISTINCT u.customer, u.store
FROM Purchase u, Purchase x, Product y
WHERE u.store = x.store
    AND y.price > 1000
    AND x.product = y.pname
```

Notice that Purchase occurs twice. Why?

Modified query:

```
SELECT DISTINCT u.customer, u.store
FROM Purchase u,
  (SELECT DISTINCT x.store, y.price
   FROM  Purchase x, Product y
   WHERE x.product = y.pname) v
WHERE u.store = v.store
    AND v.price > 1000
```

Modified and unnested query:

Purchase(customer, product, store)
Product(pname, price)

StorePrice(store, price)

Further Virtual View Optimization

Retrieve all stores whose name contains ACME

```
CREATE VIEW StorePrice AS
  SELECT DISTINCT x.store, y.price
  FROM  Purchase x, Product y
  WHERE x.product = y.pname
```

```
SELECT DISTINCT v.store
FROM StorePrice v
WHERE v.store like '%ACME%'
```

Purchase(customer, product, store)
Product(pname, price)

StorePrice(store, price)

Further Virtual View Optimization

Retrieve all stores whose name contains ACME

```
CREATE VIEW  StorePrice AS
  SELECT DISTINCT x.store, y.price
  FROM  Purchase x, Product y
  WHERE x.product = y.pname
```

```
SELECT DISTINCT v.store
FROM StorePrice v
WHERE v.store like '%ACME%'
```

Modified query:

```
SELECT DISTINCT v.store
FROM
 (SELECT DISTINCT x.store, y.price
  FROM  Purchase x, Product y
  WHERE x.product = y.pname) v
WHERE v.store like '%ACME%'
```

Purchase(customer, product, store)
Product(pname, price)

StorePrice(store, price)

Further Virtual View Optimization

Retrieve all stores whose name contains ACME

```
SELECT DISTINCT x.store
FROM Purchase x, Product y
WHERE x.product = y.pname
   AND  x.store like '%ACME%'
```

We can further optimize! How?

Modified query:

```
SELECT DISTINCT v.store
FROM
  (SELECT DISTINCT x.store, y.price
   FROM  Purchase x, Product y
   WHERE x.product = y.pname) v
WHERE v.store like '%ACME%'
```

Modified and unnested query:

Purchase(customer, product, store)
Product(<u>pname</u>, price)

StorePrice(store, price)

Further Virtual View Optimization

Retrieve all stores whose name contains ACME

SELECT DISTINCT x.store
FROM Purchase x, ~~Product y~~
WHERE ~~x.product = y.pname~~
~~AND~~ x.store like ‘%ACME%’

Assuming Product.pname is a key *and* Purchase.product is a foreign key

Modified and unnested query:

Final Query

```
SELECT DISTINCT x.store
FROM Purchase x
WHERE x.store like ‘%ACME%’
```

Applications of Virtual Views

- Increased physical data independence. E.g.
  - Vertical data partitioning
  - Horizontal data partitioning
- Logical data independence. E.g.
  - Change schemas of base relations (i.e., stored tables)
- Security
  - View reveals only what the users are allowed to know

Vertical Partitioning

Resumes

<u>SSN</u>	Name	Address	Resume	Picture
234234	Mary	Huston	Clob1…	Blob1…
345345	Sue	Seattle	Clob2…	Blob2…
345343	Joan	Seattle	Clob3…	Blob3…
432432	Ann	Portland	Clob4…	Blob4…

T1

<u>SSN</u>	Name	Address
234234	Mary	Huston
345345	Sue	Seattle
. . .		

T2

<u>SSN</u>	Resume
234234	Clob1…
345345	Clob2…

T3

<u>SSN</u>	Picture
234234	Blob1…
345345	Blob2…

T2.SSN is a key *<u>and</u>* a foreign key to **T1**.SSN. Same for **T3**.SSN

T1(ssn,name,address)
T2(ssn,resume)
T3(ssn,picture)

Resumes(ssn,name,address,resume,picture)

Vertical Partitioning

```
CREATE VIEW  Resumes  AS
  SELECT  T1.ssn, T1.name, T1.address,
          T2.resume, T3.picture
  FROM    T1,T2,T3
  WHERE   T1.ssn=T2.ssn AND T1.ssn=T3.ssn
```

T1(ssn,name,address)
T2(ssn,resume)
T3(ssn,picture)

Resumes(ssn,name,address,resume,picture)

Vertical Partitioning

```
CREATE VIEW  Resumes  AS
   SELECT  T1.ssn, T1.name, T1.address,
           T2.resume, T3.picture
   FROM    T1,T2,T3
   WHERE  T1.ssn=T2.ssn AND T1.ssn=T3.ssn
```

```
SELECT address
FROM    Resumes
WHERE   name = 'Sue'
```

T1(ssn,name,address)
T2(ssn,resume)
T3(ssn,picture)

Resumes(ssn,name,address,resume,picture)

Vertical Partitioning

```
CREATE VIEW  Resumes  AS
   SELECT  T1.ssn, T1.name, T1.address,
           T2.resume, T3.picture
   FROM    T1,T2,T3
   WHERE   T1.ssn=T2.ssn AND T1.ssn=T3.ssn
```

```
SELECT address
FROM    Resumes
WHERE   name = 'Sue'
```

Modified query:

```
SELECT T1.address
FROM T1, T2, T3
WHERE T1.name = 'Sue'
    AND T1.SSN=T2.SSN
    AND T1.SSN = T3.SSN
```

T1(ssn,name,address)
T2(ssn,resume)
T3(ssn,picture)

Resumes(ssn,name,address,resume,picture)

Vertical Partitioning

```
CREATE VIEW  Resumes  AS
   SELECT  T1.ssn, T1.name, T1.address,
           T2.resume, T3.picture
   FROM    T1,T2,T3
   WHERE   T1.ssn=T2.ssn AND T1.ssn=T3.ssn
```

```
SELECT address
FROM   Resumes
WHERE  name = 'Sue'
```

Modified query:

SELECT T1.address
FROM T1, ~~T2, T3~~
WHERE T1.name = 'Sue'
~~AND T1.SSN=T2.SSN~~
~~AND T1.SSN = T3.SSN~~

Final query:

```
SELECT T1.address
FROM T1
WHERE T1.name = 'Sue'
```

Vertical Partitioning Applications

1. Advantages
   – Speeds up queries that touch only a small fraction of columns
   – Single column can be compressed effectively, reducing disk I/O

2. Disadvantages
   – Updates are expensive!
   – Need many joins to access many columns
   – Repeated key columns add overhead

Hot trend today for data analytics: e.g., Vertica startup acquired by HP
They use a highly-tuned column-oriented data store AND engine

Horizontal Partitioning

Customers

SSN	Name	City
234234	Mary	Houston
345345	Sue	Seattle
345343	Joan	Seattle
234234	Ann	Portland
--	Frank	Calgary
--	Jean	Montreal

CustomersInHouston

SSN	Name	City
234234	Mary	Houston

CustomersInSeattle

SSN	Name	City
345345	Sue	Seattle
345343	Joan	Seattle

. . . . .

CustomersInHouston(ssn,name,city)
CustomersInSeattle(ssn,name,city)
. . . . .

Customers(ssn,name,city)

Horizontal Partitioning

```
CREATE VIEW  Customers  AS
    CustomersInHouston
       UNION ALL
    CustomersInSeattle
       UNION ALL
     . . .
```

CustomersInHouston(ssn,name,city)
CustomersInSeattle(ssn,name,city)
. . . . .

Customers(ssn,name,city)

Horizontal Partitioning

```
SELECT name
FROM   Customers
WHERE  city = 'Seattle'
```

Which tables are inspected by the system ?

CustomersInHouston(ssn,name,city)
CustomersInSeattle(ssn,name,city)
. . . . .

Customers(ssn,name,city)

Horizontal Partitioning

```
SELECT name
FROM    Customers
WHERE   city = 'Seattle'
```

Which tables are inspected by the system ?

All tables!
The systems doesn't know that CustomersInSeattle.city = 'Seattle'

CustomersInHouston(ssn,name,city)
CustomersInSeattle(ssn,name,city)
. . . . .

Customers(ssn,name,city)

Horizontal Partitioning

Better: remove CustomerInHuston.city etc

```
CREATE VIEW  Customers  AS
    (SELECT SSN, name, 'Houston' as city
     FROM CustomersInHouston)
       UNION ALL
    (SELECT SSN, name, 'Seattle' as city
     FROM CustomersInSeattle)
       UNION ALL
     . . .
```

CustomersInHouston(<u>ssn</u>,name,city)
CustomersInSeattle(<u>ssn</u>,name,city)
. . . . .

Customers(<u>ssn</u>,name,city)

Horizontal Partitioning

```
SELECT  name
FROM    Customers
WHERE   city = 'Seattle'
```

↓

```
SELECT name
FROM   CustomersInSeattle
```

Horizontal Partitioning Applications

- Performance optimization
  - Especially for data warehousing
  - E.g. one partition per month
  - E.g. archived applications and active applications
- Distributed and parallel databases
- Data integration