CSE 4125: Distributed Database Systems
Chapter – 6
(Part – A)

Optimization of Access Strategies.

Pre-requisites

- Knowledge of Chapter 2

Topics to be discussed -

- Database Profiles
- Estimating Database Profiles

Database Profiles

What are Database Profiles?

- *Statistical information of the database.*

- *Necessity:*
  - –To perform sequence of operations, relations must be transmitted over the network.
  - –It is important to estimate the size of the results to minimize the data transfers.
  - –We need the statistical information (**database profile)** to estimate.

Information in Database Profiles

For a relation $R(A_1, A_2, ..., A_n)$ with fragments $R_1, R_2, ..., R_r$ the database profile contains following information.

- **card (R_i):** number of tuples of R_i.
- **size (A_i):** size or length (i.e. number of bytes) of attribute A_i.
- **size (R_i):** sum of the size of all attributes of Ri.

Information in Database Profiles

For a relation $R(A_1, A_2, ..., A_n)$ with fragments $R_1, R_2, ..., R_r$ the database profile contains following information.

- **val (A_i [R_i]):** number of distinct values appearing for attribute A_i of R_i.
- **dom(A_i):** domain of an attribute.
- **site (R_i):** allocated site of the fragment R_i.

Database Profiles (example)

- card ($DEPT_1$) = 10
- site($DEPT_1$) = 2

	deptnum	name	area	mgrnum
size	2	15	1	7
val	10	10	2	10

- Q: size (R_i) = ?

Database Profiles (example)

- card ($DEPT_1$) = 10
- site($DEPT_1$) = 2

	deptnum	name	area	mgrnum
size	2	15	1	7
val	10	10	2	10

- Q: size (R_i) = 25

Exercise 1

Given three relations R, S, T. We want to perform the following query Q.

Where, $\mathbf{Q = PJ_a\,(R\ UN\ S)}$

Card (R) = 3000 **Card (S) = 1000** **Card (T) = 4000**

	a	b	c	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	c	b	d
Size	6	7	2	10
Val	1000	20	500	100

	a	m	n
Size	6	5	4
Val	4000	10	5

Assume that , the result of (R UN S) has no duplicate values in attribute 'a'

Is the database profile correct? Mention and rewrite the errors if any. 1

Estimating profiles of results of algebraic operations

What to Estimate?

- Estimating the profiles of results of algebraic operations.
- This information is useful for optimization.
- Assume, *R and S* are input fragments and *T* is the result.
  - We will mostly estimate ***card(T)*** and ***size(T).***
  - Example: If *card(R)* and *card(S)* is given, can we estimate *card(T)* for *T = R UN S*

Selection

$$T = SL_{A = value}\ R$$

Cardinality

$$card\ (T) = \rho \times card\ (R)$$

Here ρ is selectivity.

$$\rho = 1\ /\ val\ (\ A[R]\ )$$

Selection

$$T = SL_{A = value}\ R$$

Cardinality

$$card(T) = \rho \times card(R)$$

Here ρ is selectivity.

$$\rho = 1 / val\ (\ A[R]\ )$$

example

R

A1	*A2*
A	E
B	F
C	G
D	H

$T = SL_{A1 = B}\ R$

ρ = ?

card(T) = ?

Selection

$$T = SL_{A = value}\ R$$

Cardinality

card (T) = $\rho \times$ card (R)

Here ρ is selectivity.

$\rho = 1 / val\ (\ A[R]\ )$

example

R

A1	A2
A	E
B	F
A	G
B	H

T = $SL_{A1 = B}$ **R**

ρ = ?

card(T) = ?

***Assuming, values are homogeneously distributed*

Selection

$$T = SL_{A = value}\ R$$

Size

size (T) = ?

Selection

$$T = SL_{A = value}\ R$$

Size

$$size\ (T) = size\ (R)$$

- *Selection doesn't affect the size of relations.*

Projection

$$T = PJ_{A1} R$$

Cardinality

card (T) = ?

example

R

A1	*A2*
A	E
B	F
C	G
B	H

Projection

$$T = PJ_{A1}\, R$$

Cardinality

card (T) = val (A1[R])

example

R

A1	*A2*
A	E
B	F
C	G
B	H

Projection

$$T = PJ_{A1}\, R$$

Size

size (T) **?** size (?)

example

R

A1	*A2*
A	E
B	F
C	G
B	H

Projection

$$T = PJ_{A1}\, R$$

Size

size (T) = size (A1)

example

R

A1	*A2*
A	E
B	F
C	G
B	H

Union

$$T = R \text{ UN } S$$

Cardinality

$$\text{card}(T) \leq \text{card}(R) + \text{card}(S)$$

***Equality holds when duplicate tuples are not eliminated.*

Union

T = R UN S

Size

size (T) = size (R) = size (S)

$$\text{card (Result)} \leq \text{card (R)} + \text{card (S)}$$
$$\text{size (Result)} = \text{size (R)} = \text{size (S)}$$

Example: *R UN S*

R

A	B	C
a	1	a
b	1	b
a	1	d
b	2	f

S

A	B	C
a	1	a
a	3	f

T

B	C	D
1	a	1
3	b	1
3	c	2
1	d	4
2	a	3

Result

A	B	C
a	1	a
b	1	b
a	1	d
b	2	f
a	3	f

Difference

$$T = R \; DF \; S$$

Cardinality

$$\max(0, card(R) - card(S)) \leq card(T) \leq card(R)$$

Size

$$size(T) = size(R) = size(S)$$

$$\max(0, \text{card}(R) - \text{card}(S)) \leq \text{card}(\text{Result}) \leq \text{card}(R)$$
$$\text{size}(\text{Result}) = \text{size}(R) = \text{size}(S)$$

Example: $R \; DF \; S$

R		
A	**B**	**C**
a	1	a
b	1	b
a	1	d
b	2	f

S		
A	**B**	**C**
a	1	a
a	3	f

T		
B	**C**	**D**
1	a	1
3	b	1
3	c	2
1	d	4
2	a	3

Result		
A	**B**	**C**
b	1	b
a	1	d
b	2	f

Cartesian Product

$$T = R \; CP \; S$$

Cardinality

$$card\ (T) = card\ (R) \times card\ (S)$$

Size

$$size\ (T) = size\ (R) + size\ (S)$$

$$\text{card (Result)} = \text{card (R)} \times \text{card (S)}$$
$$\text{size (Result)} = \text{size (R)} + \text{size (S)}$$

Example: $R \; CP \; S$

R

A	B	C
a	1	a
b	1	b
a	1	d
b	2	f

S

A	B	C
a	1	a
a	3	f

T

B	C	D
1	a	1
3	b	1
3	c	2
1	d	4
2	a	3

Result

R.A	R.B	R.C	S.A	S.B	S.C
a	1	a	a	1	a
b	1	b	a	1	a
a	1	d	a	1	a
b	2	f	a	1	a
a	1	a	a	3	f
b	1	b	a	3	f
a	1	d	a	3	f
b	2	f	a	3	f

Join

$$T = R \; JN_{R.A = S.B} \; S$$

Cardinality

$$\text{card}(T) = \text{selectivity} \times \text{card}(R \; CP \; S)$$

$$= \rho \times (\text{card}(R) \times \text{card}(S))$$

Join

$$T = R \text{ JN}_{R.A = S.B} S$$

Cardinality

$$\text{card}(T) = \text{selectivity} \times \text{card}(R \; CP \; S)$$
$$= \rho \times (\text{card}(R) \times \text{card}(S))$$
$$= 1 / val\,(A[R]) \times \text{card}(R) \times \text{card}(S)$$
$$= (\text{card}(R) \times \text{card}(S)) / val\,(A[R])$$
$$= (\text{card}(R) \times \text{card}(S)) / val\,(B[S])$$

Join

$$T = R \text{ JN}_{R.A = S.B} S$$

Size

$$\text{size}(T) = \text{size}(R) + \text{size}(S)$$

$$\text{card (Result)} = (\text{card(R)} \times \text{card(T)}) \,/\, val\,(\text{C}[R])$$
$$\text{size (Result)} = \text{size (R)} + \text{size (S)}$$

Example: $R \; JN_{R.C=T.C} \; T$

R

A	B	C
a	1	a
b	1	b
a	1	d
b	2	f

S

A	B	C
a	1	a
a	3	f

T

B	C	D
1	a	1
3	b	1
3	c	2
1	d	4
2	a	3

Result

A	R.B	R.C	T.B	T.C	D
a	1	a	1	a	1
a	1	a	2	a	3
b	1	b	3	b	1
a	1	d	1	d	4

Semi-Join

$$T = R \; SJ_{R.A = S.B} \; S$$

Cardinality

$$\text{card}(T) = \text{selectivity} \times \text{card}(R)$$
$$= \rho \times \text{card}(R)$$

Here,

$$\rho = val(A[S]) / val(\text{dom}(A))$$

Size

The size of the result of a semi-join is the same size of its first operand.

Estimating Example 1

Given three relations R, S, T. We want to perform the following query Q.

Where, $\mathbf{Q} = (PJ_a\,(R\ UN\ S))\ CP\ (T\ JN_{a=a}\ S)$

Card (R) = 3000 **Card (S) = 1000** **Card (T) = 4000**

	a	b	c	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	b	c	d
Size	6	7	2	10
Val	1000	20	500	100

	a	m	n
Size	6	5	4
Val	4000	10	5

Assume that , the result of (R UN S) has no duplicate values in attribute 'a'

Estimate the cardinality of the result of query Q. Indicate necessary formulas applied to estimate. 4

Given Query: $(\boldsymbol{PJ_a}\ (\boldsymbol{R\ UN\ S}))\ \boldsymbol{CP}\ (\boldsymbol{T\ JN_{a=a}S})$

Let,

X = **R UN S**

Y = $\boldsymbol{PJ_a}$ (X)

Z = **T** $\boldsymbol{JN_{a=a}}$ **S**

F = Y $\boldsymbol{CP}$ Z

Card(X) = Card (R) + Card(S) = 3000 + 1000
= 4000

Card(Y) = val(a[X]) = 4000

Card (Z) = ρ × (card(T) × card(S))
= (1 / val (a[T])) * Card(T) * card(S)
= (1 / 4000) * 4000 * 1000
= 1000

Card(F) = Card(Y) * card(Z)
= 4000 * 1000 = 4000000

Estimating Example 2

Given three relations R, S, T. We want to perform the following query Q.

Where, $\mathbf{Q} = (PJ_a\,(R\ UN\ S))\ CP\ (T\ JN_{a=a}\ S)$

Card (R) = 3000 **Card (S) = 1000** **Card (T) = 4000**

	a	b	c	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	b	c	d
Size	6	7	2	10
Val	1000	20	500	100

	a	m	n
Size	6	5	4
Val	4000	10	5

Assume that , the result of (R UN S) has no duplicate values in attribute 'a'

Estimate the size of the result of query Q. Indicate necessary formulas applied to estimate. 4

Given Query: $(\boldsymbol{PJ_a}\ (\boldsymbol{R\ UN\ S}))\ \boldsymbol{CP}\ (\boldsymbol{T\ JN_{a=a}S})$

Let,

X = **R UN S**

Y = $\boldsymbol{PJ_a}$ (X)

Z = **T** $\boldsymbol{JN_{a=a}}$ **S**

F = Y $\boldsymbol{CP}$ Z

size(X) = size (R)
= size(a) + size(b) + size(c) + size(d)
= 6+7+2+10 = 25 bytes

size(Y) = size(a) = 6 bytes

size(Z) = size(T) + size(S)
= (6+5+4) + (6+7+2+10)
= 40 bytes

size(F) = size(Y) + size(Z)

= 6 + 40 = 46 bytes = 46 * 8 bits = 368 bits

Exercise 2

Given three relations R, S, T. We want to perform the following query Q.

Where, **Q** = $PJ_a\,(R\ UN\ S)$

Card (R) = 3000 **Card (S) = 1000** **Card (T) = 4000**

	a	b	c	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	c	b	d
Size	6	7	2	10
Val	1000	20	500	100

	a	m	n
Size	6	5	4
Val	4000	10	5

Assume that , the result of (R UN S) has no duplicate values in attribute 'a'

Is the database profile correct? Mention and rewrite the errors if any. 1

Estimate the cardinality of the result of query Q. Indicate necessary formulas applied to estimate. 4

What is the output of $size(R\ \ SJ_{a=a}\ T)$? 1

Exercise 3

Given three relations R, S, T. We want to perform the following query Q.

Where, $Q = (PJ_a(R\ UN\ S))\ JN_{a=a}\ T$

Card (R) = 3000 **Card (S) = 1000** **Card (T) = 4000**

	a	b	c	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	b	c	d
Size	6	7	2	10
Val	1000	20	500	100

	a	m	n
Size	6	5	4
Val	4000	10	5

Assume that , the result of (R UN S) has no duplicate values in attribute 'a'

Estimate the cardinality & size of query Q. Indicate necessary formulas applied to estimate. 4