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.

\square *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.

- \triangleright card ($\mathbf{R_i}$): number of tuples of R_i .
- > size (A_i) : size or length (i.e. number of bytes) of attribute A_i .
- \triangleright size (\mathbf{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.

- \triangleright val (A_i [R_i]): number of distinct values appearing for attribute A_i of R_i .
- \rightarrow dom(A_i): domain of an attribute.
- \triangleright site ($\mathbf{R_i}$): allocated site of the fragment R_i .

Database Profiles (example)

- card $(DEPT_1) = 10$ Number of rows
 - $site(DEPT_1) = 2$ Dept 1 fragment is located at Site 2

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

• Q: size $(R_i) = ?$ 2+15+1+7 = 25 byte

Database Profiles (example)

- card (DEPT₁) = 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,
$$Q = PJ_a(R UN S)$$

$$Card(R) = 3000$$

$$Card(S) = 1000$$

$$Card(T) = 4000$$

	a	ъ	C	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	С	Ъ	đ
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.
- \triangleright 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 $T = R \ UN \ S$

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

Cardinality

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

Here ρ is selectivity.

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

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

Cardinality

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

Here ρ is selectivity.

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

example

R

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

$$\rho = ?$$

$$card(T) = ?$$

A1	A2
Α	E
В	F
С	G
D	Н

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

Cardinality

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

Here ρ is selectivity. $\rho = 1 / val(A[R])$

resumple
$$R$$

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

$$\rho = ?$$

$$card(T) = ?$$

$$R$$

$$A1 \quad A2$$

$$A \quad E$$

$$B \quad F$$

$$A \quad G$$

$$B \quad H$$

**Assuming, values are homogeneously distributed

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

Size

size
$$(T) = ?$$

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

Size

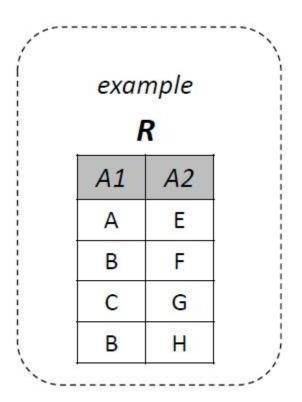
$$size(T) = size(R)$$

> Selection doesn't affect the size of relations.

$$\left\{ \mathbf{T} = \mathbf{PJ}_{A1} R \right\}$$

Cardinality

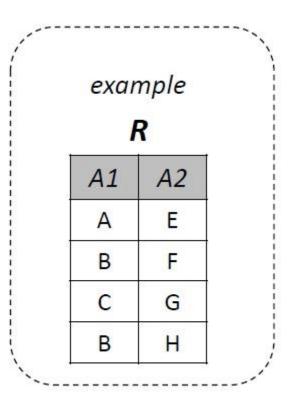
card(T) = ?



$$T = PJ_{A1}R$$

Cardinality

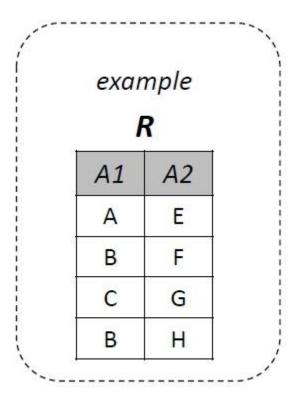
card(T) = val(A1[R])



$$T = PJ_{AI}R$$

Size

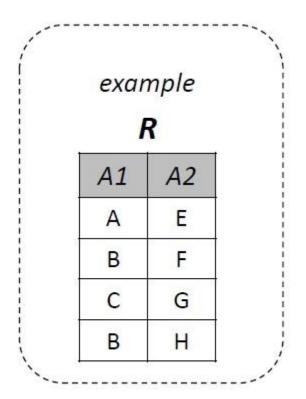
size (T) ? size (?)



$$\left\{ \mathbf{T} = \mathbf{PJ}_{A1} R \right\}$$

Size

size
$$(T)$$
 = size $(A1)$



Union

$$T = R UN S$$

Cardinality

$$card(T) \le card(R) + card(S)$$

**Equality holds when duplicate tuples are not eliminated.

Union

$$T = R UN S$$

Size

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

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

Example: R UNS

R				
В	С			
1	а			
1	b			
1	d			
2	f			
	1 1			

S				
Α	В	С		
а	1	а		
а	3	f		

	T	
В	С	D
1	а	1
3	b	1
3	С	2
1	d	4
2	а	3

Result			
Α	В	С	
а	1	а	
b	1	b	
а	1	d	
b	2	f	
а	3	f	

Difference

$$T = R DF S$$

Cardinality

$$\max (0, \operatorname{card} (R) - \operatorname{card} (S)) \le \operatorname{card} (T) \le \operatorname{card} (R)$$

Size

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

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

Example: *R DF S*

R			
В	С		
1	а		
1	b		
1	d		
2	f		
	1 1 1		

S				T	
А	В	С	В	С	D
а	1	a	1	а	1
а	3	f	3	b	1
			3	С	2
			1	d	4
			2	а	3

Result				
Α	В	С		
b	1	b		
а	1	d		
b	2	f		

Cartesian Product

$$T = R CP S$$

Cardinality

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

Size

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

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

Example: R CPS

R			
A	В	С	
а	1	а	
b	1	b	
а	1	d	
b	2	f	

S		
Α	В	С
а	1	а
а	3	f

	T	
В	С	D
1	а	1
3	b	1
3	С	2
1	d	4
2	а	3

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

Join

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

Cardinality

card (T) = selectivity × card(R
$$CP$$
 S)
= ρ × (card(R) × card(S))

Join

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

Cardinality

```
card (T) = selectivity × card(R CP S)

= \rho × (card(R) × card(S))

= 1 / val (A[R]) × card(R) × card(S)

= (card(R) × card(S)) / val (A[R])

= (card(R) × card(S)) / val (B[S])
```

Join

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

Size

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

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

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

R			
Α	В	С	
а	1	а	
b	1	b	
а	1	d	
b	2	f	

S		
Α	В	С
а	1	a
а	3	f

Т	
С	D
а	1
b	1
С	2
d	4
а	3
	C a b c

	Result					
Α	R.B	R.C	T.B	T.C	D	
а	1	а	1	а	1	
а	1	а	2	a	3	
b	1	b	3	b	1	
а	1	d	1	d	4	

Semi-Join

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

Cardinality

card (T) = selectivity × card(R)
=
$$\rho$$
 × card(R)

Here,

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

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

	a	Ъ	С	đ
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: $(PJ_a(RUNS))CP(TJN_{a=a}S)$

Let,

$$X = R UN S$$

 $Y = PJ_a(X)$
 $Z = T JN_{a=a} S$
 $F = Y CP Z$

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

= 4000

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

Card (Z) =
$$\rho \times (\text{card}(T) \times \text{card}(S))$$

= $(1 / \text{val } (a[T])) * \text{Card}(T) * \text{card}(S)$
= $(1 / 4000) * 4000 * 1000$
= 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,
$$Q = (PJ_a(R UN S)) CP(T JN_{a=a} S)$$

$$Card(R) = 3000$$

$$Card(S) = 1000$$

$$Card(T) = 4000$$

	a	ь	С	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	b	С	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: $(PJ_a(RUNS))CP(TJN_{a=a}S)$

Let,

$$X = R UN S$$

 $Y = PJ_a(X)$
 $Z = T JN_{a=a} S$
 $F = Y CP Z$

$$size(X) = size(R)$$

$$= size(a) + size(b) + size(c) + size(d)$$

$$= 6+7+2+10 = 25 \text{ bytes}$$

$$size(Y) = size(a) = 6 \text{ bytes}$$

$$size(Z) = size(T) + size(S)$$

$$= (6+5+4) + (6+7+2+10)$$

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

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

= 40 bytes

Exercise 2

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

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

$$Card(R) = 3000$$

$$Card(S) = 1000$$

$$Card(T) = 4000$$

	a	ь	С	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	С	Ъ	đ
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.

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

1

Exercise 3

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

Where,
$$Q = (PJ_a(RUNS))JN_{a=a}T$$

$$Card(R) = 3000$$

$$Card(S) = 1000$$

$$Card(T) = 4000$$

	a	ь	С	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	Ъ	С	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