

CSCI317 Database Performance Tuning

Clustering

Dr Janusz R. Getta

School of Computing and Information Technology -
University of Wollongong

Clustering

Outline

Clustering ? What is it ?

Sample clustered data structures

Creating a cluster

Loading relational tables into a cluster

Clustering – costs versus benefits

Optimal clustering

Suboptimal clustering

Clustering ? What is it ?

Cluster is a group of tables that share the same data blocks because they share the same columns and are frequently used together

Clusters are transparent to query languages

Clusters are logically and physically dependent of the data in the associated tables

Once created, a **cluster** is automatically maintained and used by a database system

Performance related observations

Retrieval performance of **clustered tables** may be better than retrieval performance of non-clustered tables

Presence of **clusters** decreases performance of **UPDATE, DELETE, and INSERT** operations

Clustering

Outline

Clustering ? What is it ?

Sample clustered data structures

Creating a cluster

Loading relational tables into a cluster

Clustering – costs versus benefits

Optimal clustering

Suboptimal clustering

Sample clustered data structures

Consider the relational tables **EMP** and **DEPT** with information about the employees and departments the employees are located at

EMP

EMPNO	ENAME	DEPTID
932	Peter	CS
654	Michael	IT
345	Mary	IT
286	Joan	IT
507	John	CS

DEPT

DEPTID	DNAME	LOC
CS	COMP. SCI.	3
IT	INF. TECH.	6

EMP-DEPT CLUSTER

<u>DEPTID</u>		
CS		
<u>DNAME</u>	<u>LOC</u>	
COMP. SCI.	3	
<u>EMPNO</u>	<u>ENAME</u>	
932	Peter	
507	John	
<u>DEPTID</u>		
IT		
<u>DNAME</u>	<u>LOC</u>	
INF. TECH.	6	
<u>EMPNO</u>	<u>ENAME</u>	
654	Michael	
345	Mary	
286	Joan	

EMP-DEPT cluster contains the rows from a relational table **DEPT** grouped by **DEPTID** and joined with the rows from a relational table **EMP**

Clustering

Outline

Clustering ? What is it ?

Sample clustered data structures

Creating a cluster

Loading relational tables into a cluster

Clustering – costs versus benefits

Optimal clustering

Suboptimal clustering

Creating a cluster

A **cluster** can be created in the following way

```
CREATE CLUSTER COUNTRY_CLST  
  (CO_ID NUMBER(4) )  
  INDEX  
  PCTFREE 0;
```

Creating a cluster

A **cluster** must have an index created on its key

```
CREATE INDEX COUNTRY_IDX ON CLUSTER COUNTRY_CLST;
```

Creating an index on a cluster

Clustering

Outline

Clustering ? What is it ?

Sample clustered data structures

Creating a cluster

Loading relational tables into a cluster

Clustering – costs versus benefits

Optimal clustering

Suboptimal clustering

Loading the relational tables into a cluster

A relational table **COUNTRY** can be created in a **cluster COUNTRY_CLST** in the following way

```
CREATE TABLE COUNTRY(  
  CO_ID          NUMERIC(4)      NOT NULL,  
  CO_NAME        VARCHAR(50)     NOT NULL,  
  ...            ...            ...  
  CONSTRAINT COUNTRY_PKEY PRIMARY KEY (CO_ID),  
  CONSTRAINT COUNTRY_CHECK1 CHECK(CO_ID > 0) )  
CLUSTER COUNTRY_CLST(CO_ID);
```

Creating a relational table in a cluster

A relational table **ADDRESS** can be created in a **cluster COUNTRY_CLST** in the following way

```
CREATE TABLE ADDRESS(  
  ADDR_ID        NUMERIC(10)    NOT NULL,  
  ADDR_CO_ID     NUMERIC(4)     NOT NULL,  
  ...            ...            ...  
  CONSTRAINT ADDRESS_PKEY PRIMARY KEY (ADDR_ID),  
  CONSTRAINT ADDRESS_FKEY FOREIGN KEY (ADDR_CO_ID)  
    REFERENCES COUNTRY(CO_ID),  
  CONSTRAINT ADDRESS_CHECK1 CHECK(ADDR_ID > 0) )  
CLUSTER COUNTRY_CLST(ADDR_CO_ID);
```

Creating a relational table in a cluster

Clustering

Outline

Clustering ? What is it ?

Sample clustered data structures

Creating a cluster

Loading relational tables into a cluster

Clustering – costs versus benefits

Optimal clustering

Suboptimal clustering

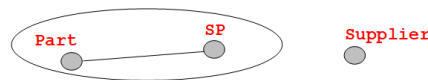
Clustering costs versus benefits

Consider the relational tables

```
PART(pnum, pname, price)
SP(pnum, snum, qty)
SUPPLIER(snum, sname, address)
```

Sample relational tables to be clustered

Variant 1: Clustering of **PART** and **SP** over **pnum**



Benefits: no need to process join of relational tables **PART** and **SP**

Costs: instead of join operation a cluster over **pnum** must be read

$$(\text{cost}(\text{PART JOIN SP}) - \text{cost}(\text{cluster}(\text{PART UNION SP}))) * \text{frequency}(\text{PART JOIN SP})$$

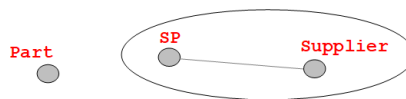
Clustering costs versus benefits

Consider the relational tables

```
PART(pnum, pname, price)
SP(pnum, snum, qty)
SUPPLIER(snum, sname, address)
```

Sample relational tables to be clustered

Variant 2: Clustering of **SUPPLIER** and **SP** over **snum**



Benefits: no need to process join of relational tables **SUPPLIER** and **SP**

Costs: instead of join operation a cluster over **pnum** must be read

$$\begin{aligned} & (\text{cost}(\text{SUPPLIER JOIN SP}) - \text{cost}(\text{cluster}(\text{SUPPLIER UNION} \\ & \text{SP}))) * \\ & \text{frequency}(\text{SUPPLIER JOIN SP}) \end{aligned}$$

Clustering costs versus benefits

Consider the relational tables

```
PART(pnum, pname, price)
SP(pnum, snum, qty)
SUPPLIER(snum, sname, address)
```

Sample relational tables to be clustered

$\text{size}(\text{SP}(\text{pnum}, \text{snum}, \text{qty})) = 50 \text{ blocks}$

$\text{size}(\text{SUPPLIER}(\text{snum}, \text{sname}, \text{address})) = 100 \text{ blocks}$

$\text{cost}(\text{PART JOIN SP}) = 7000 \text{ read blocks, frequency} = 10 \text{ times per day}$

$\text{cost}(\text{SUPPLIER JOIN SP}) = 2500 \text{ read blocks, frequency} = 30 \text{ times per day}$

Variant 1

- $\text{Savings} = (7000 - (300+50)) * 10 = 66500 \text{ read blocks}$

Variant 2

- $\text{Savings} = (2500 - (100+50)) * 30 = 70500 \text{ read blocks}$

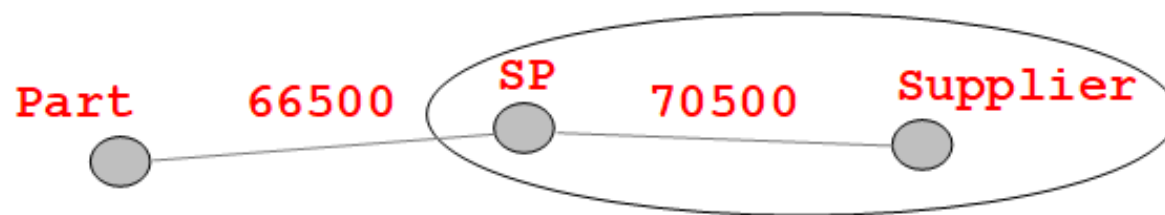
Clustering costs versus benefits

Variant 1

- Savings = $(7000 - (300 + 50)) * 10 = 66500$ read blocks

Variant 2

- Savings = $(2500 - (100 + 50)) * 30 = 70500$ read blocks



Clustering

Outline

Clustering ? What is it ?

Sample clustered data structures

Creating a cluster

Loading relational tables into a cluster

Clustering – costs versus benefits

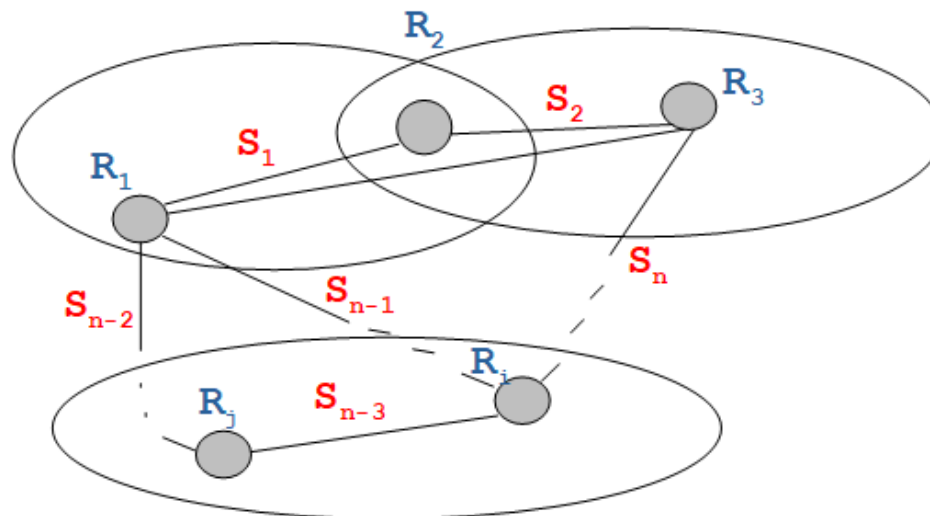
Optimal clustering

Suboptimal clustering

Optimal clustering

Problem

- Given a binary graph (clustering graph) representing all possible clustering variants together with evaluation of benefits for each one of them
- Find the smallest set of variants V that maximizes the savings



Relational tables: $R_1, R_2, \dots, R_j$

Benefits: $S_1, S_2, \dots, S_n$

Clustering

Outline

Clustering ? What is it ?

Sample clustered data structures

Creating a cluster

Loading relational tables into a cluster

Clustering – costs versus benefits

Optimal clustering

Suboptimal clustering

Suboptimal clustering

Algorithm

Suboptimal clustering algorithm

Make a set of clustering variants V empty

repeat

Find in a clustering graph a variant $V_{\max}$ that maximises savings;

Add $V_{\max}$ to V ;

Remove from a clustering graph an edge that represents a variant $V_{\max}$
and all edges that represent variants inconsistent with $V_{\max}$;

until clustering graph has no edges;

Suboptimal clustering

Assume that database consists of the following relational tables:

- R size 100 blocks
- S size 50 blocks
- T size 200 blocks
- U size 80 blocks
- V size 50 blocks

Assume that the tables

- R and S are joined on average 10 times per day
- S and T are joined on average 5 times per day
- T and U are joined on average 10 times per day
- T and V are joined on average 24 times per day

Suboptimal clustering

Assume that join of

- R and S needs 200 read block operations
- S and T needs 300 read block operations
- T and U needs 300 read block operations
- T and V needs 400 read block operations

Then, the benefits form clustering of

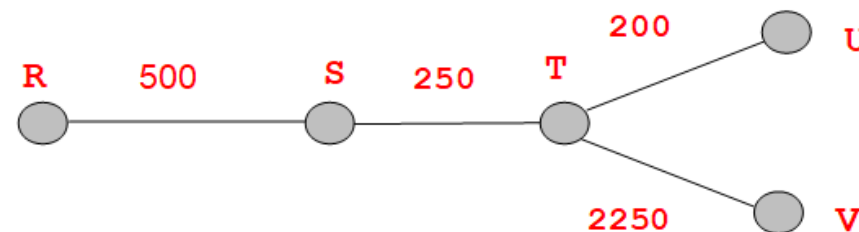
- R and S are equal to $10 * 200 - 10 * (100 + 50) = 500$ reads per day
- S and T are equal to $5 * 300 - 5 * (50 + 200) = 250$ reads per day
- T and U are equal to $10 * 300 - 10 * (200 + 80) = 200$ reads per day
- T and V are equal to $15 * 400 - 15 * (200 + 50) = 2250$ reads per day

Suboptimal clustering

Then, the benefits form clustering of

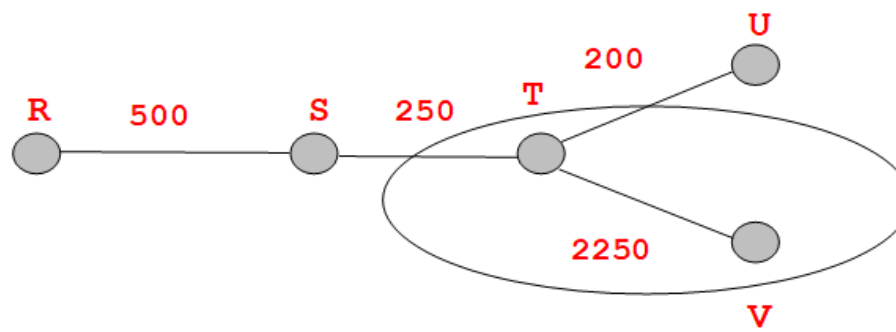
- R and S are equal to $10 * 200 - 10 * (100 + 50) = 500$ reads per day
- S and T are equal to $5 * 300 - 5 * (50 + 200) = 250$ reads per day
- T and U are equal to $10 * 300 - 10 * (200 + 80) = 200$ reads per day
- T and V are equal to $15 * 400 - 15 * (80 + 50) = 2250$ reads per day

Clustering graph

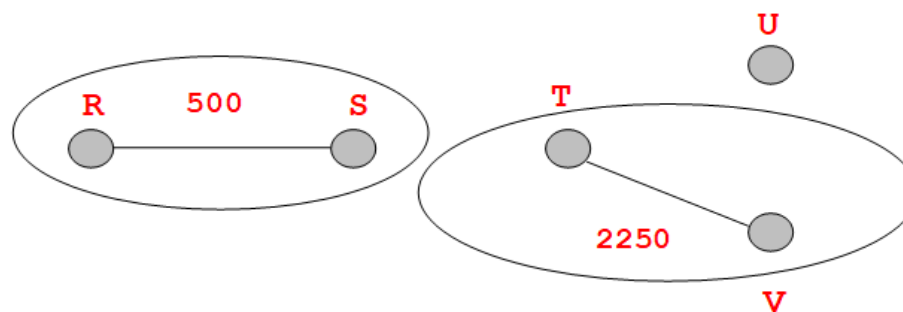


Suboptimal clustering

Cluster **T** and **V**



Cluster **R** and **S**



References

[Cookbook, How to cluster relational tables ?](#)

Ramakrishnan R., J. Gehrke Database Management Systems, chapter 20.4.1