CSCI317 – DATABASE PERFORMANCE TUNING

Tutorial

Clustering Relational Tables

Sionggo Japit

sjapit@uow.edu.au

17 August 2022

What is a cluster?

- 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
- 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

Question

Sample Clustering Structures

Sample clustering structures

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

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

 CSCI315 Database Design & Implementation (SIM, January 2012) 3

Clustering Relational Tables

- With 3 relations (tables), there are 3 possible clusters.
- With 4 relations (tables), there are 6 possible clusters.
- With 5 relations (tables), there are 10 possible clusters.
- With 6 relations (tables), how many possible clusters are there?

$$\frac{n(n-1)}{2} \text{ possible clusters.}$$

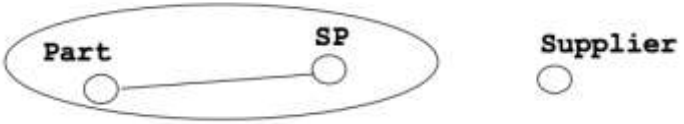
Clustering Relational Tables

Consider the relational tables

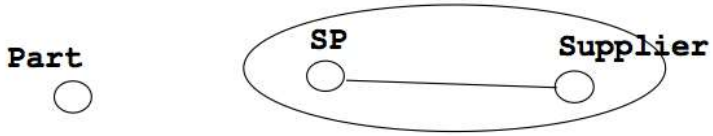
- Part(p#, pname, price)
- SP(p#, s#, qty)
- Supplier(s#, sname, address)

Two possible clusters are:

1. Part and SP over p#



2. SP and Supplier over s#



Clustering Relational Tables

- Retrieval performance of clustered tables may be better than retrieval performance of non-clustered tables.
- Possible saving from clustering tables:

[(Cost of joining the two relations) -
(Cost of exhaustively search on the UNION of the two relations)] ´
(Frequency of the joining the two relations)

Clustering Relational Tables

Example:

- Consider the relational tables

  Part(p#, pname, price) occupies 300 blocks

  SP(p#, s#, qty) occupies 50 blocks

  Supplier(s#, sname, address) occupies 100 blocks

- Joining the relational tables Part and SP needs to read 7000 blocks, and it is done on average 10 timer per day.
- Joining the relational tables Supplier and SP needs to read 2500 blocks, and it is done on average 30 times per day.
- What is the best (optimal) cluster can be formed?

Clustering Relational Tables

- Saving from clustering Part and SP over p#:

  = (7000 – (300 + 50)) * 10 = 66,500 read blocks

- Saving from clustering Supplier and SP over s#:

  = (2500 – (100 + 50)) * 30 = 70,500 read blocks

  Thus we would form the cluster Supplier and SP over s#.

Clustering Relational Tables

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;

Clustering Relational Tables

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 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 15 times per day

Clustering Relational Tables

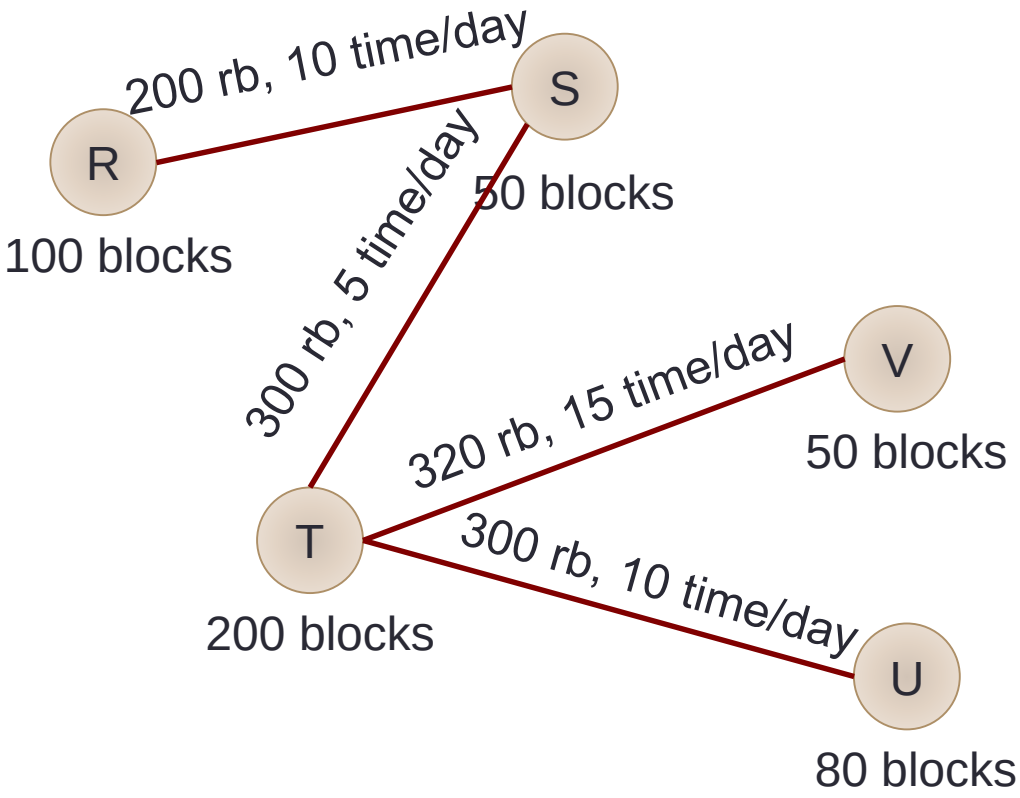
Assume that join of:

- 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 320 read block operations

Clustering Relational Tables

What are the clusters can be formed to obtain optimal saving?

Clustering Relational Tables



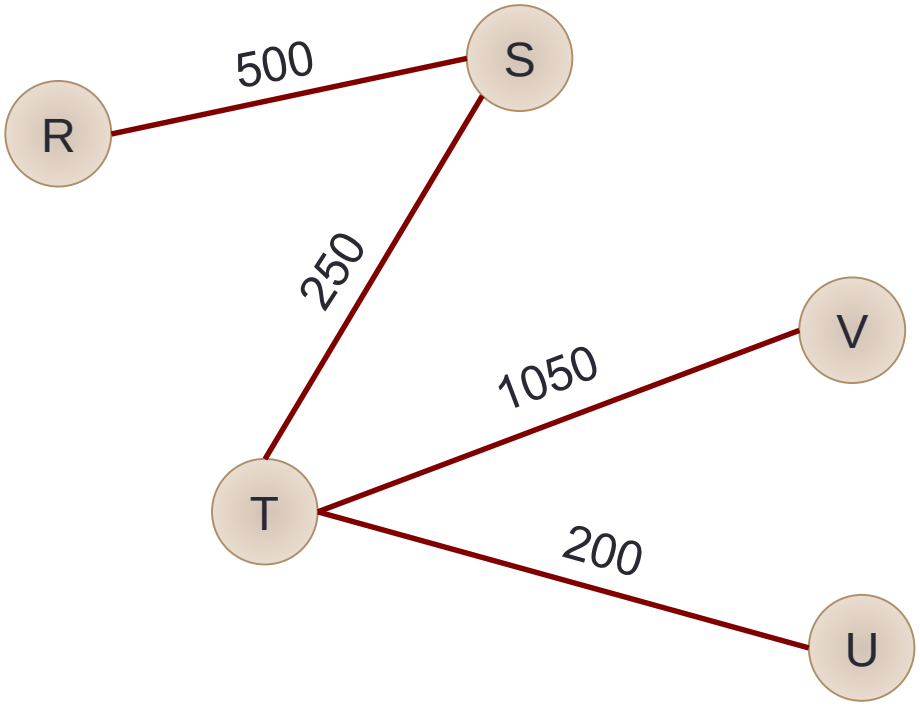
$$S_{(R,S)} = [(10 \times 200) - 10 \times (100 + 50)] = 500 \quad reads / day$$

$$S_{(S,T)} = [(5 \times 300) - 5 \times (200 + 50)] = 250 \quad reads / day$$

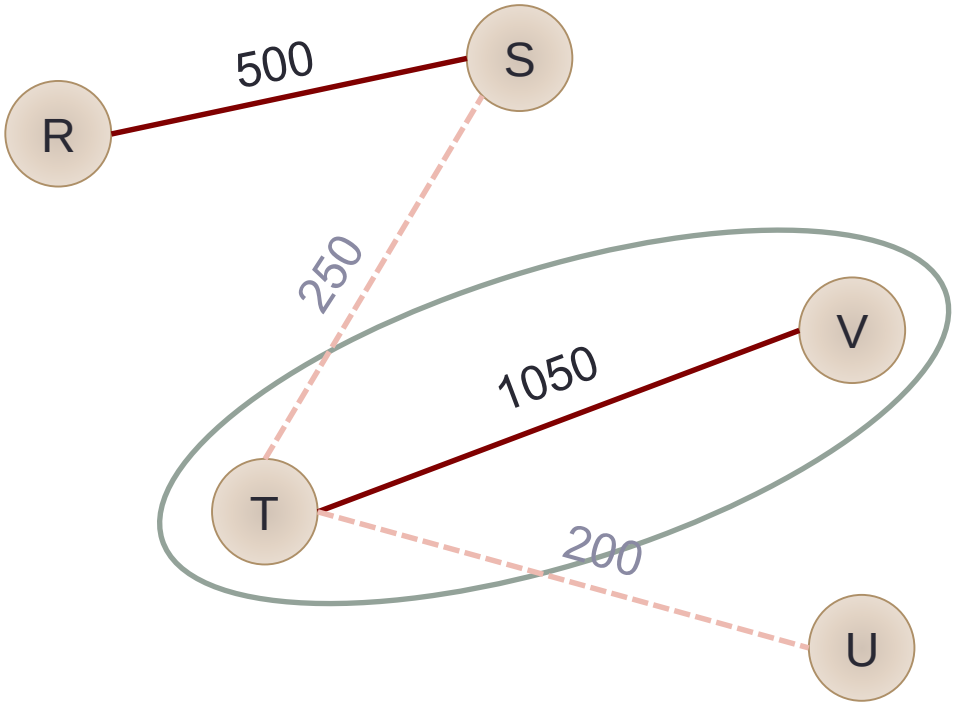
$$S_{(T,V)} = [(15 \times 320) - 15 \times (200 + 50)] = 1050 \quad reads / day$$

$$S_{(T,U)} = [(10 \times 300) - 10 \times (200 + 80)] = 200 \quad reads / day$$

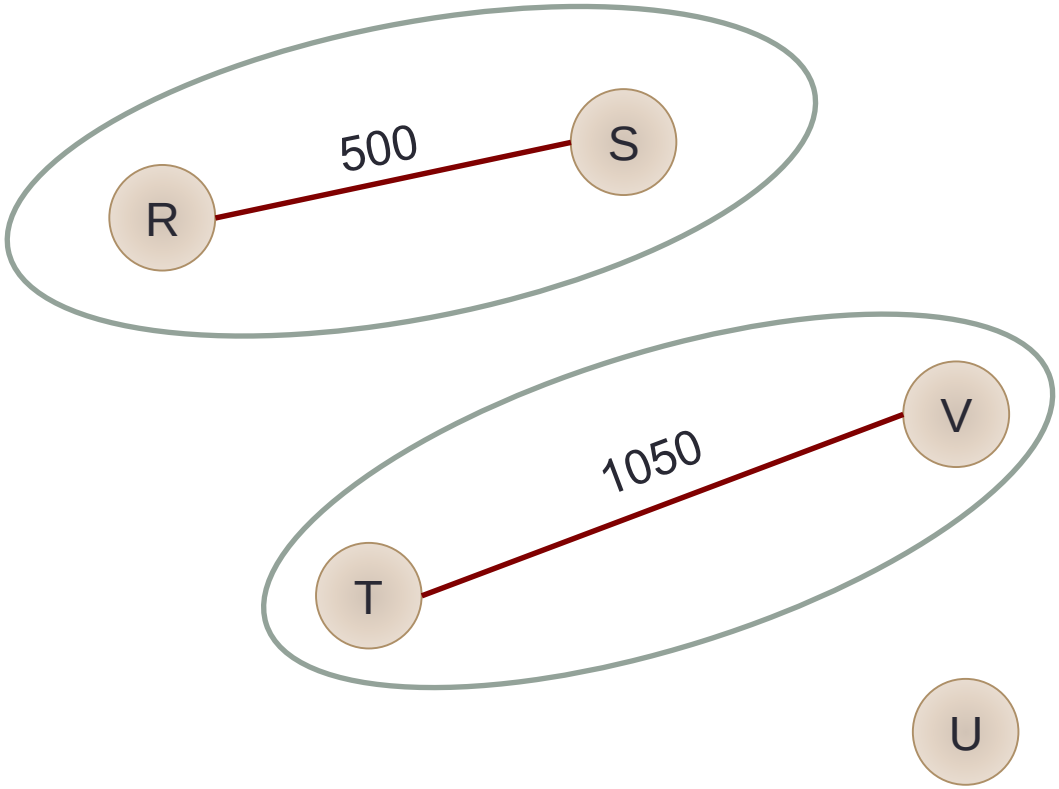
Clustering Relational Tables



Clustering Relational Tables



Clustering Relational Tables



Clustering Relational Tables