

Advanced Database Systems

Spring 2025

Lecture #11:

Joins

R&G: Chapter 14

1

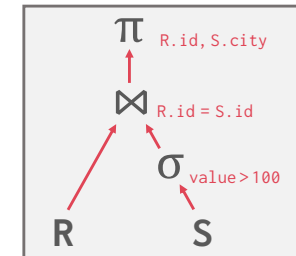
JOIN OPERATOR

4

For a tuple $r \in R$ and a tuple $s \in S$ that match on join attributes, concatenate r and s together into a new tuple

```
SELECT R.id, S.city
FROM R, S
WHERE R.id = S.id
AND S.value > 100
```

Subsequent operators in the query plan never need to go back to the base tables to get more data



4

JOINS: OVERVIEW

5

Joins are among the most **expensive** operations

of joins often used as a measure of query complexity

Join of 10s of tables common in enterprise apps

Naïve implementation: $R \bowtie_c S \equiv \sigma_c(R \times S)$

Enumerate the cross product, then filter using the join condition

Inefficient because the cross product is large

Three classes of join algorithms:

Nested loops

Sort-merge

Hash

} No particular algorithm works well in all scenarios

5

I/O COST ANALYSIS

6

Assume:

Table R has M pages and m tuples in total

Table S has N pages and n tuples in total

```
SELECT R.id, S.city
FROM R, S
WHERE R.id = S.id
AND S.value > 100
```

Cost Metric: # of I/Os to compute join

Ignore output costs (same for all join algorithms)

Ignore CPU costs

6

SIMPLE NESTED LOOPS JOIN



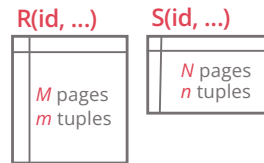
```
foreach tuple r ∈ R: ← Outer table
  foreach tuple s ∈ S: ← Inner table
    emit if r and s match
```

Why is this algorithm bad?

For every tuple in **R**, it scans **S** once

Terrible if **S** does not fit in memory

Cost: $M + (m \cdot N)$



7

SIMPLE NESTED LOOPS JOIN



Example database:

$M = 1000, m = 100,000$

$N = 500, n = 40,000$

Cost analysis:

$M + (m \cdot N) = 1000 + (100,000 \cdot 500) = 50,001,000$ I/Os

At 0.1ms per I/O, total time ≈ 1.4 hours

What if smaller table (**S**) is used as the outer table?

$N + (n \cdot M) = 500 + (40,000 \cdot 1000) = 40,000,500$ I/Os

At 0.1ms per I/O, total time ≈ 1.1 hours

8

SIMPLE NESTED LOOPS JOIN



SNLJ (but with page fetches written out explicitly)

```
foreach page PR ∈ R:
  foreach tuple r ∈ PR:
    foreach page PS ∈ S:      flip loops
      foreach tuple s ∈ PS:
        emit if r and s match
```

Can we do better?

We scan **S** for every tuple in **R**,

... but we had to load an entire page of **R** into memory to get that tuple!

Instead of finding the tuples in **S** that match a tuple in **R**,

... do the check for all tuples in a page in **R** at once

9

PAGE NESTED LOOPS JOIN

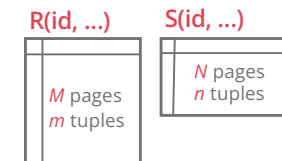
10

```
foreach page pR ∈ R:
  foreach page pS ∈ S:
    foreach tuple r ∈ pR:
      foreach tuple s ∈ pS:
        emit if r and s match
```

This algorithm makes fewer disk accesses

For every page in **R**, it scans **S** once

Cost: $M + (M \cdot N)$



10

PAGE NESTED LOOPS JOIN

11

Example database:

$M = 1000, m = 100,000$

$N = 500, n = 40,000$

Which one should be the outer table?

The smaller table in terms of # of pages

Cost analysis:

$N + (M \cdot N) = 500 + (1000 \cdot 500) = 500,500$ I/Os

At 0.1ms per I/O, total time \approx 50 seconds

How many memory buffers are needed?

Just 3: one for each input, one for output

11

BLOCK NESTED LOOPS JOIN

12

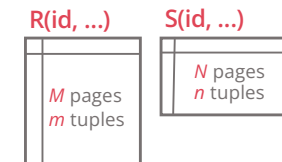
```
foreach  $B-2$  block  $b_R \in R$ :
  foreach block  $b_S \in S$ :
    foreach tuple  $r \in b_R$ :
      foreach tuple  $s \in b_S$ :
        emit if  $r$  and  $s$  match
```

What if we have B buffers available?

$B-2$ buffers for scanning the outer table

1 buffer for scanning the inner table

1 buffer for storing the output



12

BLOCK NESTED LOOPS JOIN

13

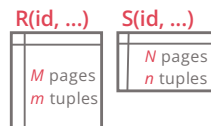
```
foreach  $B-2$  block  $b_R \in R$ :
  foreach block  $b_S \in S$ :
    foreach tuple  $r \in b_R$ :
      foreach tuple  $s \in b_S$ :
        emit if  $r$  and  $s$  match
```

Cost: $M + \lceil M / (B-2) \rceil \cdot N$

If the outer relation (R) fits in memory ($M \leq B-2$)

Cost: $M + N = 1000 + 500 = 1500$ I/Os (optimal cost)

At 0.1ms per I/O, total time \approx 0.15 seconds



13

INDEX NESTED LOOPS JOIN

14

Why do simple nested loops joins suck?

For each tuple in the outer table, we have to do a sequential scan to check for a match in the inner table

Can we accelerate the join using an index?

Use an index to find inner tuple matches

We could use an existing index or even build one on the fly

The index must match the join condition

14

INDEX NESTED LOOPS JOIN

15

```
foreach tuple r ∈ R:
  foreach tuple s ∈ Index(ri = sj)
    emit if r and s match
```

Cost: $M + m \cdot$ (cost to find all matching S tuples)

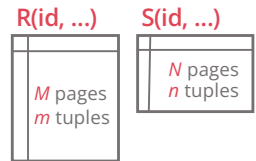
Index access cost per R tuple:

B+ tree: 2-4 I/Os to reach a leaf + fetch matching S tuples

Clustered: $M + m \cdot$ (Search + # matching pages)

Unclustered: $M + m \cdot$ (Search + up to # matching tuples)

Hash index: 1-2 I/Os to reach the target bucket



15

INDEX NESTED LOOPS JOIN

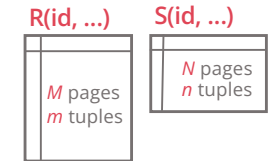
16

```
foreach tuple r ∈ R:
  foreach tuple s ∈ Index(ri = sj)
    emit if r and s match
```

Cost: $M + m \cdot$ (cost to find all matching S tuples)

The cost depends on the size of the join result

Using an index pays off if the join is **selective**



16

RECAP: NESTED LOOPS JOINS

17

Pick the smaller table as the outer table

Buffer as much of the outer table in memory as possible

Loop over the inner table

Allows arbitrary join conditions

Or use an index over the inner table

Only if matches the join condition

17

SORT-MERGE JOIN

18

Requires equality predicate

Equi-joins & natural joins

Phase #1: Sort

Sort both tables on the join key(s)

E.g. by using the external merge sort

Input might already be sorted... why?

Phase #2: Merge

Scan the two sorted tables in parallel and emit matching tuples

18

SORT-MERGE JOIN

19

```
sort R,S on join key A
r ← position of first tuple in Rsorted
s ← position of first tuple in Ssorted
while r ≠ EOF and s ≠ EOF:
  if r.A > s.A:
    advance s
  else if r.A < s.A:
    advance r
  else if r.A = s.A:
    emit (r,s)
    advance s
```

} assumes no duplicates in R
(the merge phase could be easily extended to support duplicates)

19

SORT-MERGE JOIN

20

id	name
600	Daniel
200	Michael
100	Alice
300	Bob
500	Carrol
700	Lucia
400	John

↑
Sort!

id	value	city
100	2222	Edinburgh
500	7777	Edinburgh
400	6666	London
100	9999	London
200	8888	Oxford

↑
Sort!

```
SELECT R.id, S.city
FROM R, S
WHERE R.id = S.id
AND S.value > 100
```

20

SORT-MERGE JOIN

21

id	name
100	Alice
200	Michael
300	Bob
400	John
500	Carrol
600	Daniel
700	Lucia

↑
Sort!

id	value	city
100	2222	Edinburgh
100	9999	London
200	8888	Oxford
400	6666	London
500	7777	Edinburgh

↑
Sort!

```
SELECT R.id, S.city
FROM R, S
WHERE R.id = S.id
AND S.value > 100
```

21

SORT-MERGE JOIN

22

id	name
100	Alice
200	Michael
300	Bob
400	John
500	Carrol
600	Daniel
700	Lucia

→

id	value	city
100	2222	Edinburgh
100	9999	London
200	8888	Oxford
400	6666	London
500	7777	Edinburgh

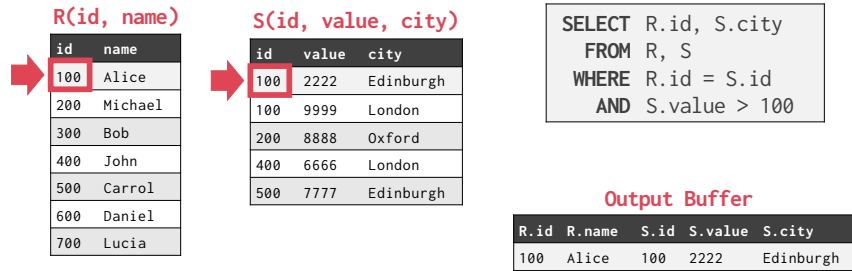
→

```
SELECT R.id, S.city
FROM R, S
WHERE R.id = S.id
AND S.value > 100
```

22

SORT-MERGE JOIN

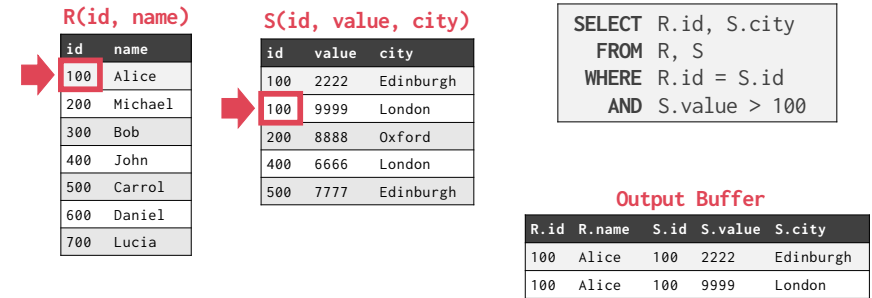
23



23

SORT-MERGE JOIN

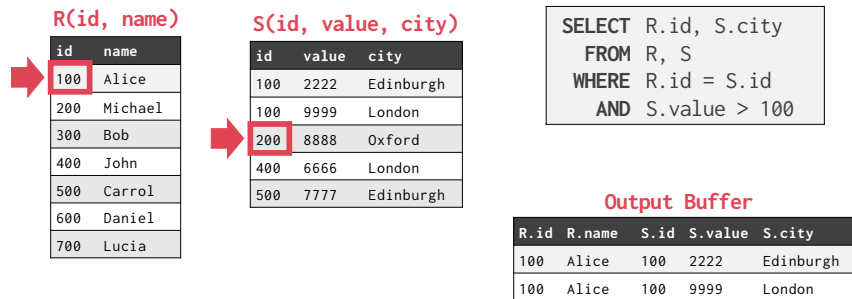
24



24

SORT-MERGE JOIN

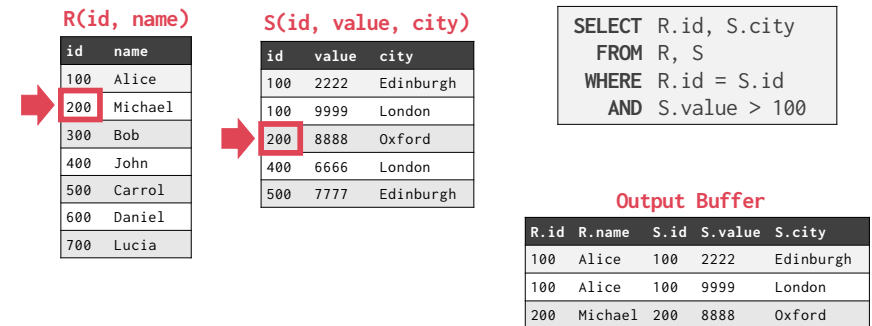
25



25

SORT-MERGE JOIN

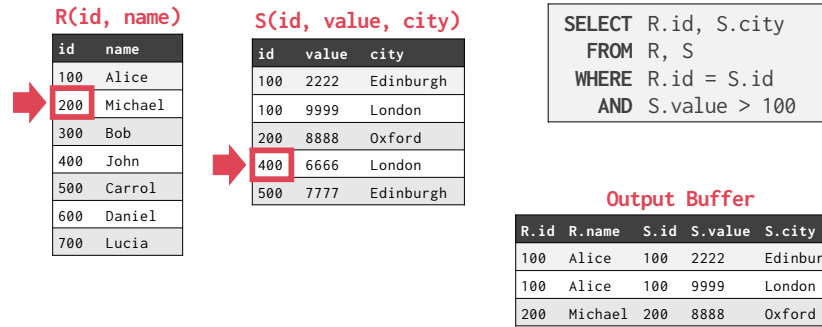
26



26

SORT-MERGE JOIN

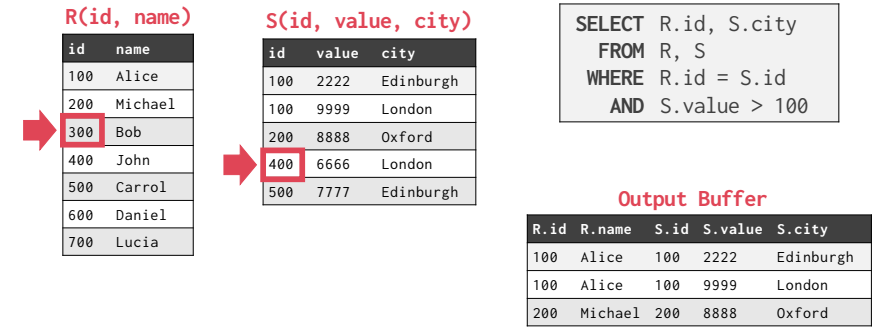
27



27

SORT-MERGE JOIN

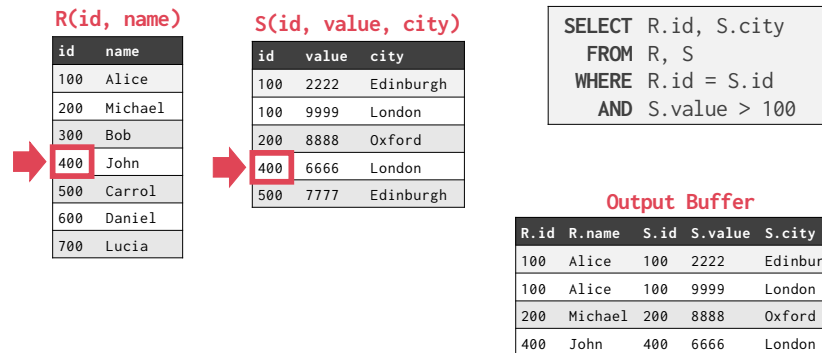
28



28

SORT-MERGE JOIN

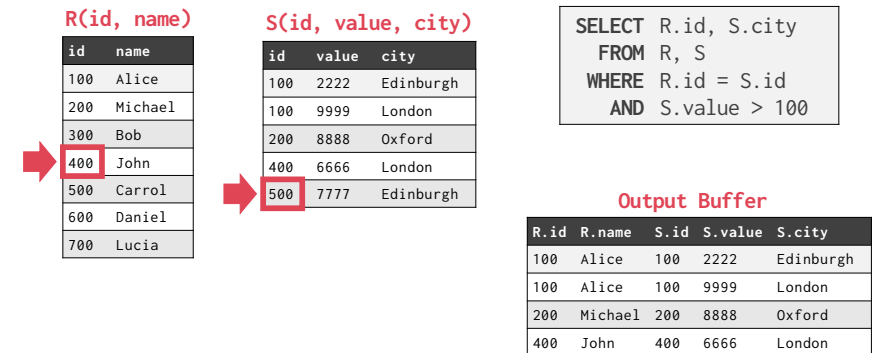
29



29

SORT-MERGE JOIN

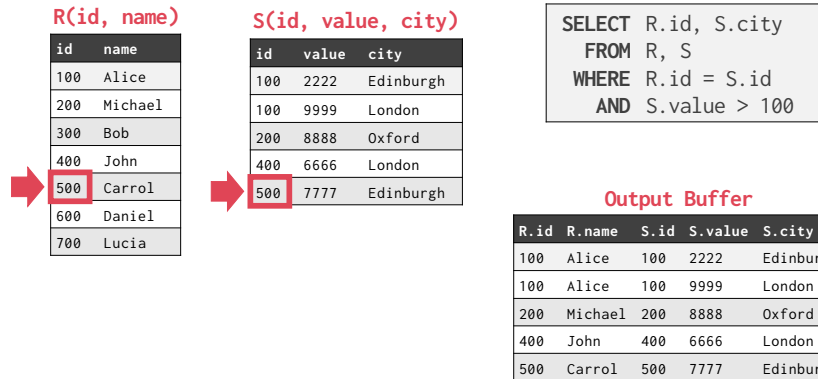
30



30

SORT-MERGE JOIN

31



31

SORT-MERGE JOIN

32

$$\text{Sort Cost (R)} = 2M \cdot (1 + \lceil \log_{B-1} \lceil M/B \rceil \rceil) \quad (= 2M \cdot \# \text{ of passes})$$

$$\text{Sort Cost (S)} = 2N \cdot (1 + \lceil \log_{B-1} \lceil N/B \rceil \rceil) \quad (= 2N \cdot \# \text{ of passes})$$

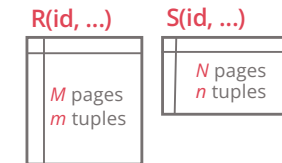
Merge Cost: $M + N$

The worst case for the merging phase is when the join attribute of all the tuples in both relations contain the same value

Sort-merge degenerates to simple nested-loops

Merge Cost: $M + m \cdot N$ (very unlikely!)

Total Cost: Sort + Merge



32

SORT-MERGE JOIN

33

Example database:

$$M = 1000, m = 100,000$$

$$N = 500, n = 40,000$$

With 100 buffer pages, both **R** and **S** can be sorted in two passes:

$$\text{Sort cost (R)} = 2 \cdot 1000 \cdot 2 = 4000 \text{ I/Os}$$

$$\text{Sort cost (S)} = 2 \cdot 500 \cdot 2 = 2000 \text{ I/Os}$$

$$\text{Merge cost} = 1000 + 500 = 1500 \text{ I/Os}$$

$$\text{Total cost} = 4000 + 2000 + 1500 = 7500 \text{ I/Os}$$

$$\text{At } 0.1 \text{ ms per I/O, total time} \approx 0.75 \text{ seconds}$$

33

SORT-MERGE JOIN REFINEMENT

34

Combine the last pass of merge-sort with the merge phase of join

Possible when the sum of # of runs in **R** and **S** in the penultimate (second-to-last) merge pass of sorting is at most $B - 1$

Example for 2-pass sort-merge join

Read **R** and write out sorted runs (pass 0)

Read **S** and write out sorted runs (pass 0)

Merge R-runs and S-runs, while finding $R \bowtie S$ matches

$$\text{Total cost} = 2M + 2N + (M + N) = 2000 + 1000 + 1500 = 4500 \text{ I/Os}$$

Eliminates one full read and write of **R** and **S**

34

WHEN IS SORT-MERGE JOIN USEFUL?

35

One or both tables are already sorted on the join key

Output must be sorted on join key (e.g., ORDER BY clause)

Typically used for **equi-joins only**

Achieves highly sequential access

Weapon of choice for very large datasets

35

BASIC IN-MEMORY HASH JOIN

36

Requires equality predicate

Phase #1: Build

Scan the outer relation and build a hash table using a hash function h on join attributes

Key: the attribute(s) that the query is joining the tables on

Value: full tuple or tuple identifier (used in column stores)

Phase #2: Probe

Scan the inner relation and use h on each tuple to jump to a location in the hash table

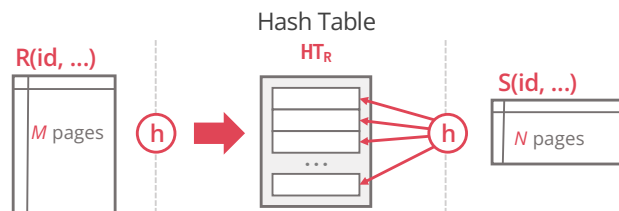
Find matching tuples there

36

BASIC IN-MEMORY HASH JOIN

37

```
build hash table  $HT_R$  for  $R$ 
foreach tuple  $s \in S$ 
  emit if  $h(s) \in HT_R$ 
```



37

HASH JOIN

38

What if both relations cannot fit in memory?

Idea: Decompose into smaller “partial joins”

If tuple $r \in R$ and tuple $s \in S$ satisfy the equi-join condition, then they have the same value for the join attributes

If that value is hashed to some value i , tuple r has to be in partition R_i and tuple s in partition S_i

Thus, R-tuples in R_i need only to be compared with S-tuples in S_i

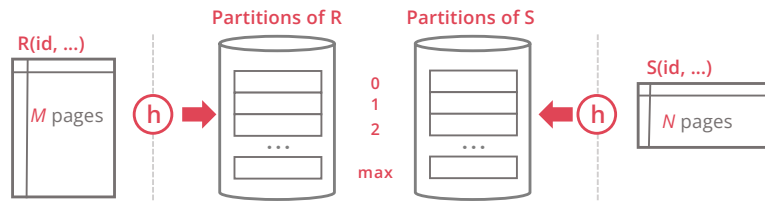
38

GRACE HASH JOIN

39

Phase #1: Partition

Partition tuples from **R** and **S** on join attribute using a hash function **h**
 Store partitions of **R** and **S** on scratch disk
 All tuples for a given join key in same partition



39

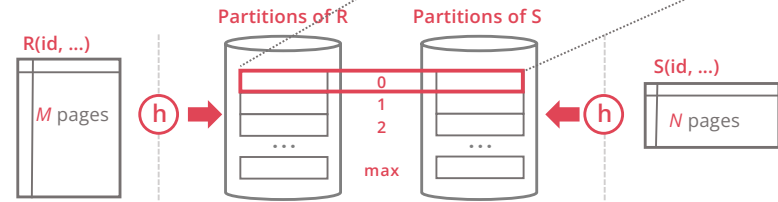
GRACE HASH JOIN

40

Phase #2: Build & Probe

Join each pair of matching partitions between **R** and **S**

build hash table $HT_{R,0}$ for R_0
 foreach tuple $s \in S_0$:
 emit if $h(s) \in HT_{R,0}$



40

GRACE HASH JOIN

41

If partitions do not fit in memory, use **recursive partitioning** with hash function $h_2 (\neq h)$ to split the partitions into chunks that will fit

In common cases, we have enough buffers to fit each pair of partitions

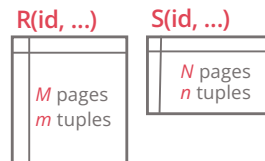
Partition Phase

Read + write both tables = $2(M + N)$ I/Os

Build & Probe Phase

Read both tables = $M + N$ I/Os

Total cost: $3(M + N)$



41

GRACE HASH JOIN

42

Example database:

$M = 1000, m = 100,000$

$N = 500, n = 40,000$

Cost Analysis:

$3 \cdot (M + N) = 3 \cdot (1000 + 500) = 4500$ I/Os

At 0.1 ms per I/O, total time ≈ 0.45 seconds

42

HASH JOIN VS. SORT-MERGE JOIN

43

Sorting pros:

- Good if input already sorted, or need output sorted
- Not sensitive to data skew or bad hash functions

Hashing pros:

- For join: # of passes depends on size of smaller relation
 - E.g. if smaller relation is $< B$, basic in-memory hashing is great
- Good if input already hashed, or need output hashed

43

JOIN ALGORITHMS: SUMMARY

44

JOIN ALGORITHM	I/O COST	TOTAL TIME
Simple Nested Loops Join	$M + (m \cdot N)$	1.4 hours
Page Nested Loops Join <small>(using 2 input and 1 output buffer)</small>	$M + (M \cdot N)$	50 seconds
Block Nested Loops Join <small>(using B memory buffers)</small>	$M + (\lceil M / (B-2) \rceil \cdot N)$	varies
Index Nested Loops Join	$M + (m \cdot \text{access cost})$	varies
Sort-Merge Join	$M + N + (\text{sort cost})$	0.75 seconds
Hash Join	$3(M + N)$	0.45 seconds
Nested Loops or Hash Join <small>(one relation fits in memory)</small>	$M + N$	0.15 seconds

44

SUMMARY

45

Nested Loops

- Works for arbitrary join condition
- Make sure to utilize memory in blocks
- Use the smaller table as the outer table

Index Nested Loops

- When you already have an index on one side
- For equi-joins mostly
- For inequality joins needs a (clustered) B+-tree index

Sort/Hash

- For equi-joins only, no index required
- Hashing better if one relation is much smaller than other
- Sorting better on non-uniform data & when results need to be sorted

No clear winners - may want to implement them all

Be sure you know the cost model for each. You will need it for query optimization!

45