Intro to Database Systems (15-445/645)

11 Join Algorithms

Carnegie Mellon University

FALL 2022

Andy Pavlo

ADMINISTRIVIA

Tuesday Oct 11th will be a pre-recorded lecture.

Homework #3 is due Sunday Oct 9th @ 11:59pm

Mid-Term Exam is Wednesday Oct 13th
→ During regular class time @ 11:50-1:10pm
→ See Piazza post for more details

Project #3 is out now:
→ Checkpoint #1: Tuesday Oct 11th @ 11:59pm
→ Checkpoint #2: Sunday Oct 23rd @ 11:59pm

WHY DO WE NEED TO JOIN?

We normalize tables in a relational database to avoid unnecessary repetition of information.

We then use the **join operator** to reconstruct the original tuples without any information loss.

JOIN ALGORITHMS

We will focus on performing binary joins (two tables) using **inner equijoin** algorithms.

→ These algorithms can be tweaked to support other joins.
→ Multi-way joins exist primarily in research literature.

In general, we want the smaller table to always be the left table ("outer table") in the query plan.

→ The optimizer will (try to) figure this out when generating the physical plan.

JOIN ALGORITHMS

We will focus on performing binary joins (two tables) using **inner equijoin** algorithms.

→ These algorithms can be tweaked to support other joins.

→ Multi-way joins exist primarily in research literature.

In general, we want the smaller table to always be the left table ("outer table") in the query plan.

→ The optimizer will (try to) figure this out when generating the physical plan.

QUERY PLAN

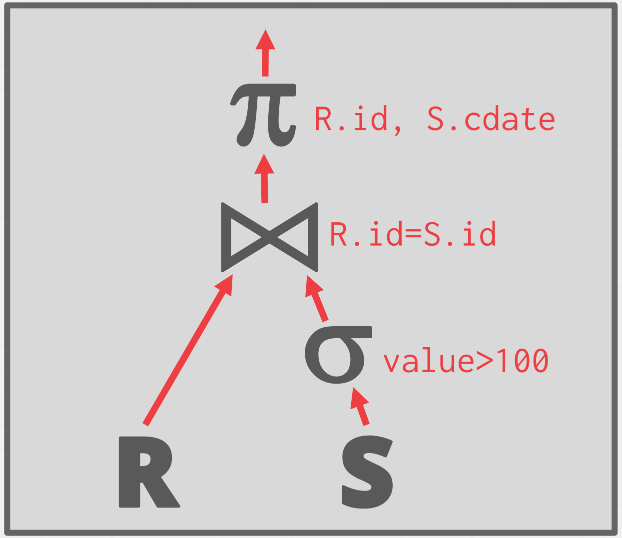
The operators are arranged in a tree.

Data flows from the leaves of the tree up towards the root.

→ We will discuss the granularity of the data movement next week.

The output of the root node is the result of the query.

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```



JOIN OPERATORS

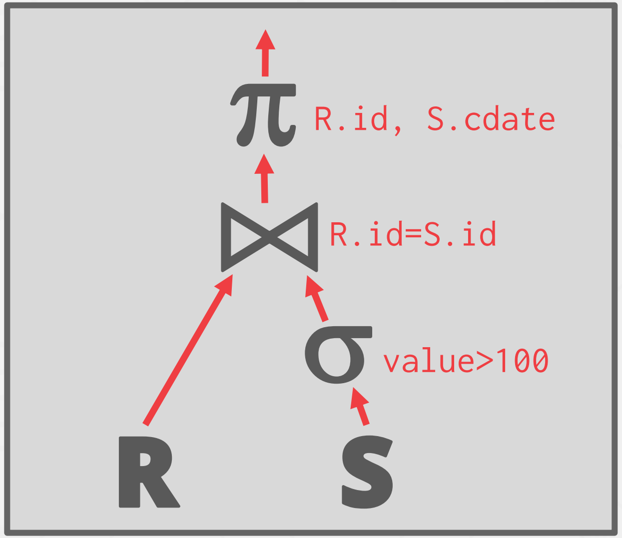
Decision #1: Output

→ What data does the join operator emit to its parent operator in the query plan tree?

Decision #2: Cost Analysis Criteria

→ How do we determine whether one join algorithm is better than another?

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```



OPERATOR OUTPUT

For tuple **r** ∈ **R** and tuple **s** ∈ **S** that match on join attributes, concatenate **r** and **s** together into a new tuple.

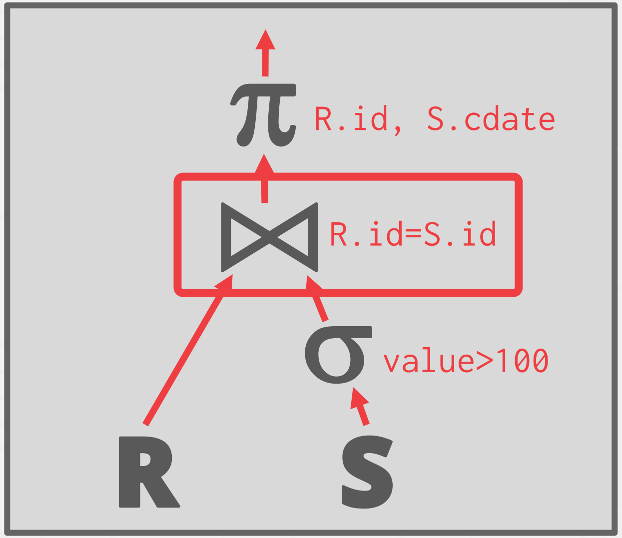
Output contents can vary:

→ Depends on processing model

→ Depends on storage model

→ Depends on data requirements in query

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```



OPERATOR OUTPUT: DATA

Early Materialization:

→ Copy the values for the attributes in outer and inner tuples into a new output tuple.

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

R(id,name)

id	name
123	abc

⋈

S(id,value,cdate)

id	value	cdate
123	1000	10/4/2022
123	2000	10/4/2022

R.id	R.name	S.id	S.value	S.cdate
123	abc	123	1000	10/4/2022
123	abc	123	2000	10/4/2022

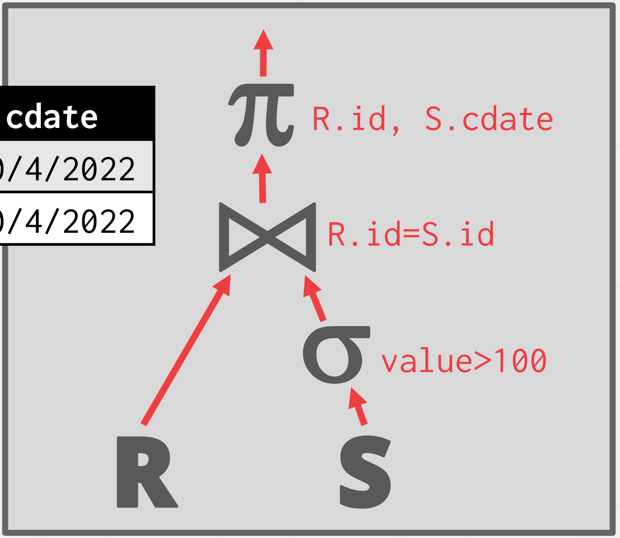
OPERATOR OUTPUT: DATA

Early Materialization:

→ Copy the values for the attributes in outer and inner tuples into a new output tuple.

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

R.id	R.name	S.id	S.value	S.cdate
123	abc	123	1000	10/4/2022
123	abc	123	2000	10/4/2022



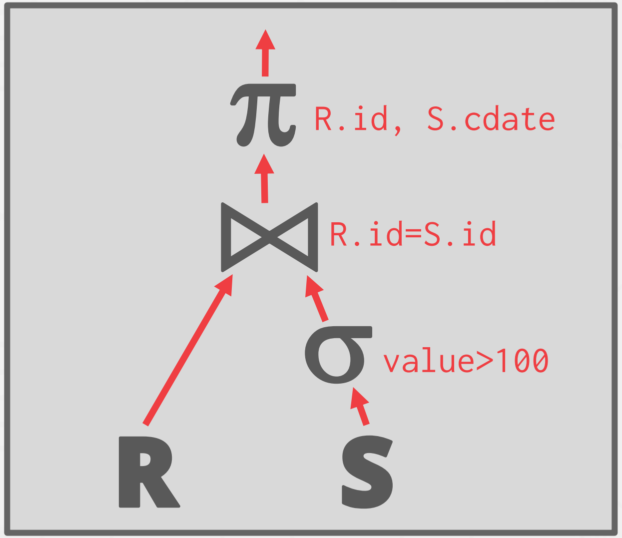
OPERATOR OUTPUT: DATA

Early Materialization:

→ Copy the values for the attributes in outer and inner tuples into a new output tuple.

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

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```



OPERATOR OUTPUT: RECORD IDS

Late Materialization:

→ Only copy the joins keys along with the Record IDs of the matching tuples.

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

R(id,name)

id	name
123	abc

⋈

S(id,value,cdate)

id	value	cdate
123	1000	10/4/2022
123	2000	10/4/2022

R.id	R.RID	S.id	S.RID
123	R.###	123	S.###
123	R.###	123	S.###

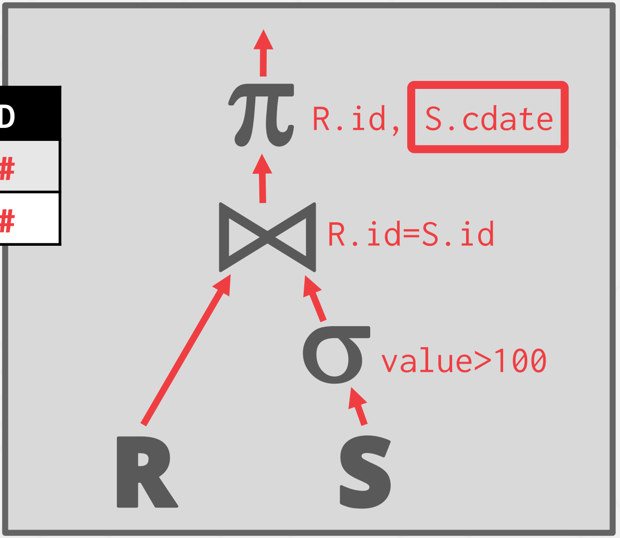
OPERATOR OUTPUT: RECORD IDS

Late Materialization:

→ Only copy the joins keys along with the Record IDs of the matching tuples.

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

R.id	R.RID	S.id	S.RID
123	R.###	123	S.###
123	R.###	123	S.###



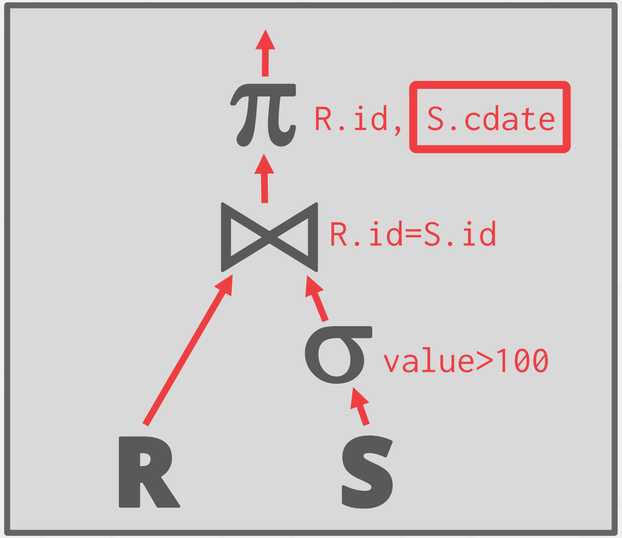
OPERATOR OUTPUT: RECORD IDS

Late Materialization:

→ Only copy the joins keys along with the Record IDs of the matching tuples.

Ideal for column stores because the DBMS does not copy data that is not needed for the query.

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```



COST ANALYSIS CRITERIA

Assume:
→ ***M*** pages in table **R**, ***m*** tuples in **R**
→ ***N*** pages in table **S**, ***n*** tuples in **S**

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

Cost Metric: # of IOs to compute join

We will ignore output costs since that depends on the data and we cannot compute that yet.

JOIN VS CROSS-PRODUCT

$R \bowtie S$ is the most common operation and thus must be carefully optimized.

$R \times S$ followed by a selection is inefficient because the cross-product is large.

There are many algorithms for reducing join cost, but no algorithm works well in all scenarios.

JOIN ALGORITHMS

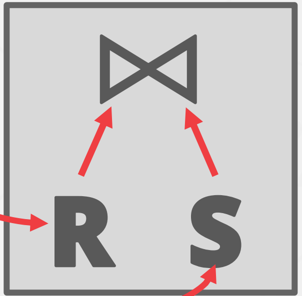
Nested Loop Join
→ Simple / Stupid
→ Block
→ Index

Sort-Merge Join

Hash Join

NESTED LOOP JOIN

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



R(id,name)

id	name
600	MethodMan
200	GZA
100	Andy
300	ODB
500	RZA
700	Ghostface
400	Raekwon

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
500	7777	10/4/2022
400	6666	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022

STUPID NESTED LOOP JOIN

Why is this algorithm stupid?
→ For every tuple in **R**, it scans **S** once

Cost: M + (m · N)

R(id,name)

M pages
m tuples

id	name
600	MethodMan
200	GZA
100	Andy
300	ODB
500	RZA
700	Ghostface
400	Raekwon

S(id,value,cdate)

N pages
n tuples

id	value	cdate
100	2222	10/4/2022
500	7777	10/4/2022
400	6666	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022

STUPID NESTED LOOP JOIN

Example database:

- Table **R**: $M = 1000$, $m = 100{,}000$
- Table **S**: $N = 500$, $n = 40{,}000$

} *4 KB pages → 6 MB*

Cost Analysis:

- $M + (m \cdot N) = 1000 + (100000 \cdot 500) =$ **50,001,000 IOs**
- At 0.1 ms/IO, Total time ≈ 1.3 hours

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

- $N + (n \cdot M) = 500 + (40000 \cdot 1000) =$ **40,000,500 IOs**
- At 0.1 ms/IO, Total time ≈ 1.1 hours

BLOCK NESTED LOOP JOIN

```
foreach block B_R ∈ R:
  foreach block B_S ∈ S:
    foreach tuple r ∈ B_R:
      foreach tuple s ∈ B_S:
        emit, if r and s match
```

R(id,name)

M pages
m tuples

id	name
600	MethodMan
200	GZA
100	Andy
300	ODB
500	RZA
700	Ghostface
400	Raekwon

S(id,value,cdate)

N pages
n tuples

id	value	cdate
100	2222	10/4/2022
500	7777	10/4/2022
400	6666	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022

BLOCK NESTED LOOP JOIN

This algorithm performs fewer disk accesses.
→ For every block in **R**, it scans **S** once.

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

R(id,name) — M pages, m tuples

id	name
600	MethodMan
200	GZA
100	Andy
300	ODB
500	RZA
700	Ghostface
400	Raekwon

S(id,value,cdate) — N pages, n tuples

id	value	cdate
100	2222	10/4/2022
500	7777	10/4/2022
400	6666	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022

BLOCK NESTED LOOP JOIN

The smaller table should be the outer table.

We determine size based on the number of pages, not the number of tuples.

R(id,name) — *M* pages, *m* tuples

id	name
600	MethodMan
200	GZA
100	Andy
300	ODB
500	RZA
700	Ghostface
400	Raekwon

S(id,value,cdate) — *N* pages, *n* tuples

id	value	cdate
100	2222	10/4/2022
500	7777	10/4/2022
400	6666	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022

BLOCK NESTED LOOP JOIN

Example database:

- → **Table R**: M = 1000, m = 100,000
- → **Table S**: N = 500, n = 40,000

Cost Analysis:

- → $M + (M \cdot N)$ = 1000 + (1000 · 500) = **501,000 IOs**
- → At 0.1 ms/IO, Total time ≈ 50 seconds

BLOCK NESTED LOOP JOIN

What if we have ***B*** buffers available?

→ Use ***B-2*** buffers for scanning the outer table.

→ Use one buffer for the inner table, one buffer for storing output.

R(id,name) — ***M*** pages, ***m*** tuples

id	name
600	MethodMan
200	GZA
100	Andy
300	ODB
500	RZA
700	Ghostface
400	Raekwon

S(id,value,cdate) — ***N*** pages, ***n*** tuples

id	value	cdate
100	2222	10/4/2022
500	7777	10/4/2022
400	6666	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022

BLOCK NESTED LOOP JOIN

```
foreach B - 2 pages pR ∈ R:
  foreach page pS ∈ S:
    foreach tuple r ∈ B - 2 pages:
      foreach tuple s ∈ pS:
        emit, if r and s match
```

R(id,name)

M pages
m tuples

id	name
600	MethodMan
200	GZA
100	Andy
300	ODB
500	RZA
700	Ghostface
400	Raekwon

S(id,value,cdate)

N pages
n tuples

id	value	cdate
100	2222	10/4/2022
500	7777	10/4/2022
400	6666	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022

BLOCK NESTED LOOP JOIN

This algorithm uses ***B*-2** buffers for scanning **R**.

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

What if the outer relation completely fits in memory ($B > M+2$)?

→ **Cost:** $M + N$ = 1000 + 500 = **1500 IOs**

→ At 0.1ms/IO, Total time ≈ 0.15 seconds

NESTED LOOP JOIN

Why is the basic nested loop join so bad?

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

We can avoid sequential scans by using an index to find inner table matches.

→ Use an existing index for the join.

INDEX NESTED LOOP JOIN

```
foreach tuple r ∈ R:
  foreach tuple s ∈ Index(r_i = s_j):
    emit, if r and s match
```

Index(S.id)

M pages
m tuples

R(id,name)

id	name
600	MethodMan
200	GZA
100	Andy
300	ODB
500	RZA
700	Ghostface
400	Raekwon

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
500	7777	10/4/2022
400	6666	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022

N pages
n tuples

INDEX NESTED LOOP JOIN

Assume the cost of each index probe is some constant C per tuple.

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

R(id,name)

M pages
m tuples

id	name
600	MethodMan
200	GZA
100	Andy
300	ODB
500	RZA
700	Ghostface
400	Raekwon

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
500	7777	10/4/2022
400	6666	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022

N pages
n tuples

Index(S.id)

NESTED LOOP JOIN: SUMMARY

Key Takeaways

→ Pick the smaller table as the outer table.
→ Buffer as much of the outer table in memory as possible.
→ Loop over the inner table (or use an index).

Algorithms

→ Simple / Stupid
→ Block
→ Index

SORT-MERGE JOIN

Phase #1: Sort

→ Sort both tables on the join key(s).
→ We can use the external merge sort algorithm that we talked about last class.

Phase #2: Merge

→ Step through the two sorted tables with cursors and emit matching tuples.
→ May need to backtrack depending on the join type.

SORT-MERGE JOIN

```
sort R,S on join keys
cursorR ← Rsorted, cursorS ← Ssorted
while cursorR and cursorS:
  if cursorR > cursorS:
    increment cursorS
  if cursorR < cursorS:
    increment cursorR
  elif cursorR and cursorS match:
    emit
    increment cursorS
```

SORT-MERGE JOIN

R(id,name)

id	name
600	MethodMan
200	GZA
100	Andy
300	ODB
500	RZA
700	Ghostface
200	GZA
400	Raekwon

Sort!

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
500	7777	10/4/2022
400	6666	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022

Sort!

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

SORT-MERGE JOIN

R(id,name)

id	name
100	Andy
200	GZA
200	GZA
300	ODB
400	Raekwon
500	RZA
600	MethodMan
700	Ghostface

Sort!

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022
400	6666	10/4/2022
500	7777	10/4/2022

Sort!

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

SORT-MERGE JOIN

R(id,name)

id	name
100	Andy
200	GZA
200	GZA
300	ODB
400	Raekwon
500	RZA
600	MethodMan
700	Ghostface

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022
400	6666	10/4/2022
500	7777	10/4/2022

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

Output Buffer

R.id	R.name	S.id	S.value	S.cdate
100	Andy	100	2222	10/4/2022

SORT-MERGE JOIN

R(id,name)

id	name
100	Andy
200	GZA
200	GZA
300	ODB
400	Raekwon
500	RZA
600	MethodMan
700	Ghostface

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022
400	6666	10/4/2022
500	7777	10/4/2022

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

Output Buffer

R.id	R.name	S.id	S.value	S.cdate
100	Andy	100	2222	10/4/2022
100	Andy	100	9999	10/4/2022

SORT-MERGE JOIN

R(id,name)

id	name
100	Andy
200	GZA
200	GZA
300	ODB
400	Raekwon
500	RZA
600	MethodMan
700	Ghostface

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022
400	6666	10/4/2022
500	7777	10/4/2022

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

Output Buffer

R.id	R.name	S.id	S.value	S.cdate
100	Andy	100	2222	10/4/2022
100	Andy	100	9999	10/4/2022
200	GZA	200	8888	10/4/2022

SORT-MERGE JOIN

R(id,name)

id	name
100	Andy
200	GZA
200	GZA
300	ODB
400	Raekwon
500	RZA
600	MethodMan
700	Ghostface

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022
400	6666	10/4/2022
500	7777	10/4/2022

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

Output Buffer

R.id	R.name	S.id	S.value	S.cdate
100	Andy	100	2222	10/4/2022
100	Andy	100	9999	10/4/2022
200	GZA	200	8888	10/4/2022

SORT-MERGE JOIN

R(id,name)

id	name
100	Andy
200	GZA
200	GZA
300	ODB
400	Raekwon
500	RZA
600	MethodMan
700	Ghostface

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022
400	6666	10/4/2022
500	7777	10/4/2022

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

Output Buffer

R.id	R.name	S.id	S.value	S.cdate
100	Andy	100	2222	10/4/2022
100	Andy	100	9999	10/4/2022
200	GZA	200	8888	10/4/2022

SORT-MERGE JOIN

R(id,name)

id	name
100	Andy
200	GZA
200	GZA
300	ODB
400	Raekwon
500	RZA
600	MethodMan
700	Ghostface

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022
400	6666	10/4/2022
500	7777	10/4/2022

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

Output Buffer

R.id	R.name	S.id	S.value	S.cdate
100	Andy	100	2222	10/4/2022
100	Andy	100	9999	10/4/2022
200	GZA	200	8888	10/4/2022
200	GZA	200	8888	10/4/2022

SORT-MERGE JOIN

R(id,name)

id	name
100	Andy
200	GZA
200	GZA
300	ODB
400	Raekwon
500	RZA
600	MethodMan
700	Ghostface

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022
400	6666	10/4/2022
500	7777	10/4/2022

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

Output Buffer

R.id	R.name	S.id	S.value	S.cdate
100	Andy	100	2222	10/4/2022
100	Andy	100	9999	10/4/2022
200	GZA	200	8888	10/4/2022
200	GZA	200	8888	10/4/2022

SORT-MERGE JOIN

R(id,name)

id	name
100	Andy
200	GZA
200	GZA
300	ODB
400	Raekwon
500	RZA
600	MethodMan
700	Ghostface

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022
400	6666	10/4/2022
500	7777	10/4/2022

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

Output Buffer

R.id	R.name	S.id	S.value	S.cdate
100	Andy	100	2222	10/4/2022
100	Andy	100	9999	10/4/2022
200	GZA	200	8888	10/4/2022
200	GZA	200	8888	10/4/2022
400	Raekwon	200	6666	10/4/2022

SORT-MERGE JOIN

R(id,name)

id	name
100	Andy
200	GZA
200	GZA
300	ODB
400	Raekwon
500	RZA
600	MethodMan
700	Ghostface

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022
400	6666	10/4/2022
500	7777	10/4/2022

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

Output Buffer

R.id	R.name	S.id	S.value	S.cdate
100	Andy	100	2222	10/4/2022
100	Andy	100	9999	10/4/2022
200	GZA	200	8888	10/4/2022
200	GZA	200	8888	10/4/2022
400	Raekwon	200	6666	10/4/2022
500	RZA	500	7777	10/4/2022

SORT-MERGE JOIN

R(id,name)

id	name
100	Andy
200	GZA
200	GZA
300	ODB
400	Raekwon
500	RZA
600	MethodMan
700	Ghostface

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022
400	6666	10/4/2022
500	7777	10/4/2022

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

Output Buffer

R.id	R.name	S.id	S.value	S.cdate
100	Andy	100	2222	10/4/2022
100	Andy	100	9999	10/4/2022
200	GZA	200	8888	10/4/2022
200	GZA	200	8888	10/4/2022
400	Raekwon	200	6666	10/4/2022
500	RZA	500	7777	10/4/2022

SORT-MERGE JOIN

R(id,name)

id	name
100	Andy
200	GZA
200	GZA
300	ODB
400	Raekwon
500	RZA
600	MethodMan
700	Ghostface

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022
400	6666	10/4/2022
500	7777	10/4/2022

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

Output Buffer

R.id	R.name	S.id	S.value	S.cdate
100	Andy	100	2222	10/4/2022
100	Andy	100	9999	10/4/2022
200	GZA	200	8888	10/4/2022
200	GZA	200	8888	10/4/2022
400	Raekwon	200	6666	10/4/2022
500	RZA	500	7777	10/4/2022

SORT-MERGE JOIN

R(id,name)

id	name
100	Andy
200	GZA
200	GZA
300	ODB
400	Raekwon
500	RZA
600	MethodMan
700	Ghostface

S(id,value,cdate)

id	value	cdate
100	2222	10/4/2022
100	9999	10/4/2022
200	8888	10/4/2022
400	6666	10/4/2022
500	7777	10/4/2022

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```

Output Buffer

R.id	R.name	S.id	S.value	S.cdate
100	Andy	100	2222	10/4/2022
100	Andy	100	9999	10/4/2022
200	GZA	200	8888	10/4/2022
200	GZA	200	8888	10/4/2022
400	Raekwon	200	6666	10/4/2022
500	RZA	500	7777	10/4/2022

SORT-MERGE JOIN

Sort Cost (R): $2M \cdot (1 + \lceil \log_{B-1} \lceil M / B \rceil \rceil)$

Sort Cost (S): $2N \cdot (1 + \lceil \log_{B-1} \lceil N / B \rceil \rceil)$

Merge Cost: $(M + N)$

Total Cost: Sort + Merge

SORT-MERGE JOIN

Example database:

- → **Table R**: M = 1000, m = 100,000
- → **Table S**: N = 500, n = 40,000

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

- → Sort Cost (**R**) = $2000 \cdot (1 + \lceil \log_{99} 1000 / 100 \rceil)$ = **4000 IOs**
- → Sort Cost (**S**) = $1000 \cdot (1 + \lceil \log_{99} 500 / 100 \rceil)$ = **2000 IOs**
- → Merge Cost = (1000 + 500) = **1500 IOs**
- → Total Cost = 4000 + 2000 + 1500 = **7500 IOs**
- → At 0.1 ms/IO, Total time ≈ 0.75 seconds

SORT-MERGE JOIN

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

Cost: (M · N) + (sort cost)

WHEN IS SORT-MERGE JOIN USEFUL?

One or both tables are already sorted on join key.

Output must be sorted on join key.

The input relations may be sorted either by an explicit sort operator, or by scanning the relation using an index on the join key.

HASH JOIN

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

If that value is hashed to some partition i, the R tuple must be in r_i and the S tuple in s_i.

Therefore, R tuples in r_i need only to be compared with S tuples in s_i.

BASIC HASH JOIN ALGORITHM

Phase #1: Build

→ Scan the outer relation and populate a hash table using the hash function h_1 on the join attributes.

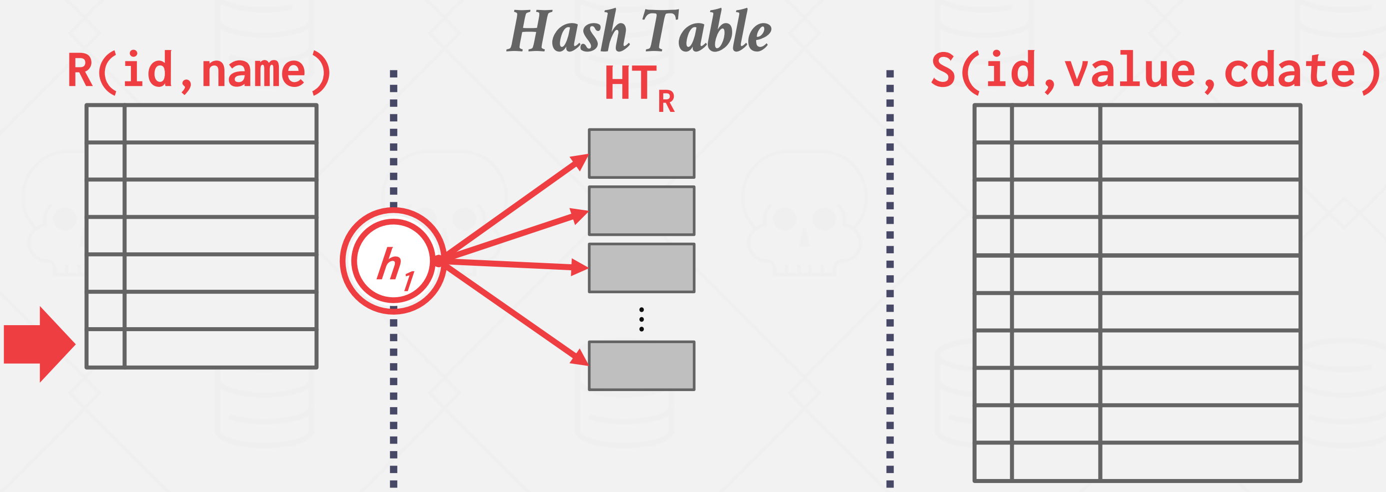
→ We can use any hash table that we discussed before but in practice linear probing works the best.

Phase #2: Probe

→ Scan the inner relation and use h_1 on each tuple to jump to a location in the hash table and find a matching tuple.

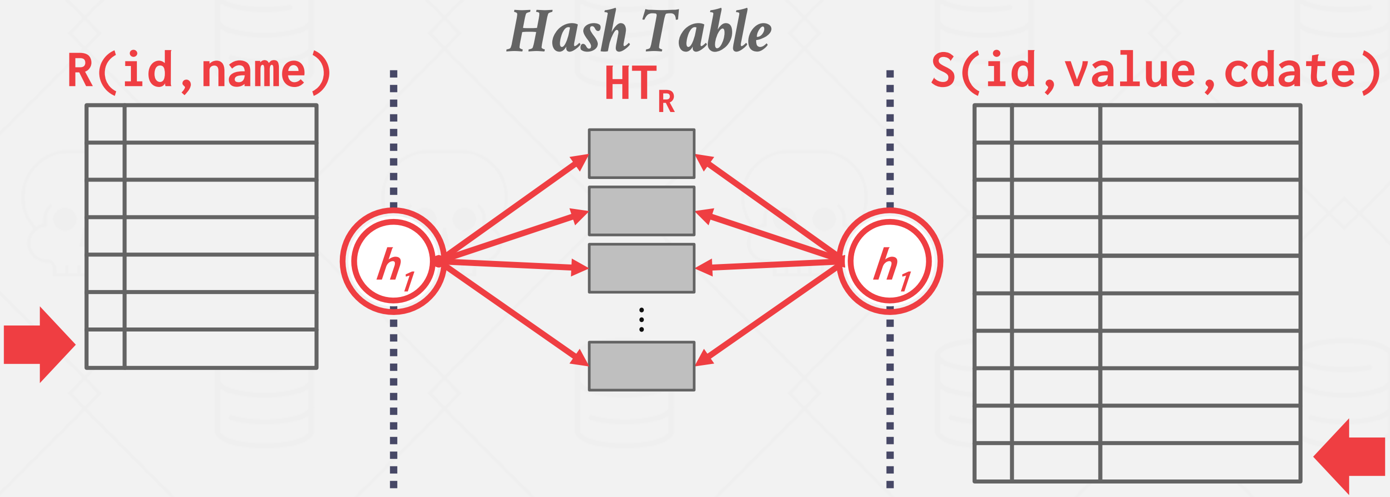
BASIC HASH JOIN ALGORITHM

```
build hash table HT_R for R
foreach tuple s ∈ S
  output, if h_1(s) ∈ HT_R
```



BASIC HASH JOIN ALGORITHM

```
build hash table HT_R for R
foreach tuple s ∈ S
  output, if h1(s) ∈ HT_R
```



HASH TABLE CONTENTS

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

- → We always need the original key to verify that we have a correct match in case of hash collisions.

Value: Varies per implementation.

- → Depends on what the operators above the join in the query plan expect as its input.
- → Early vs. Late Materialization

COST ANALYSIS

How big of a table can we hash using this approach?

→ ***B*-1** "spill partitions" in Phase #1

→ Each should be no more than ***B*** blocks big

Answer: $B \cdot (B-1)$

→ A table of N pages needs about **sqrt(*N*)** buffers

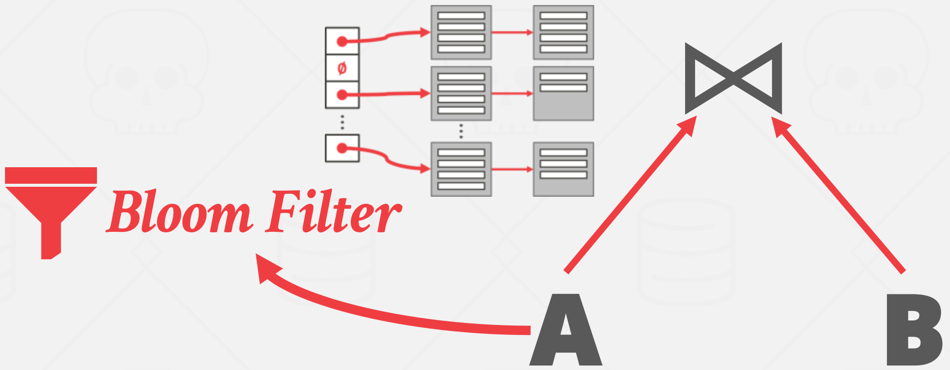
→ Assumes hash distributes records evenly. Use a "fudge factor" $f>1$ for that: we need $B \cdot \mathrm{sqrt}(f \cdot N)$

OPTIMIZATION: PROBE FILTER

Create a Bloom Filter during the build phase when the key is likely to not exist in the hash table.

→ Threads check the filter before probing the hash table. This will be faster since the filter will fit in CPU caches.

→ Sometimes called ***sideways information passing.***

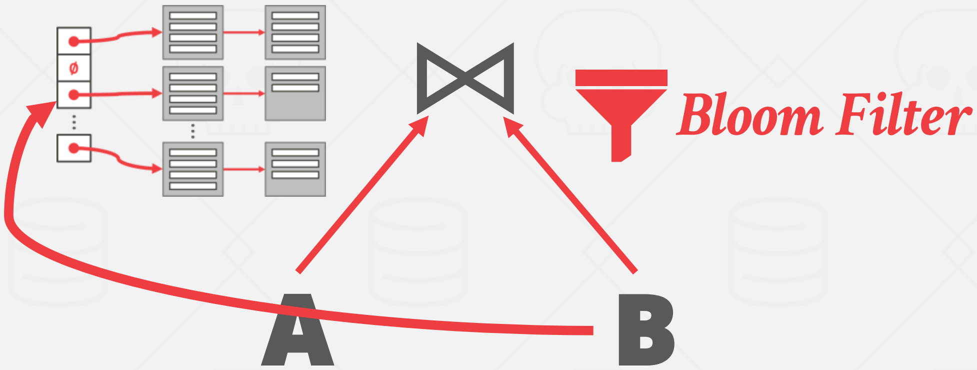


OPTIMIZATION: PROBE FILTER

Create a Bloom Filter during the build phase when the key is likely to not exist in the hash table.

→ Threads check the filter before probing the hash table. This will be faster since the filter will fit in CPU caches.

→ Sometimes called ***sideways information passing.***



BLOOM FILTERS

Probabilistic data structure (bitmap) that answers set membership queries.

→ False negatives will never occur.
→ False positives can sometimes occur.
→ See Bloom Filter Calculator.

Insert(x):

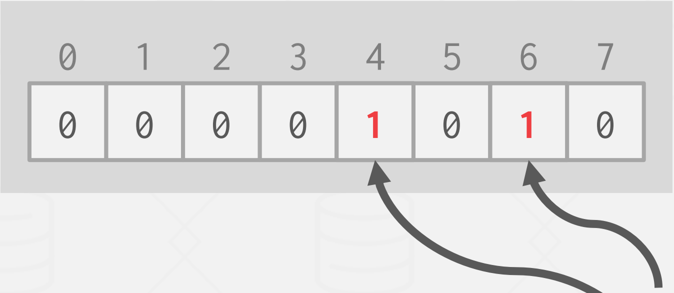
→ Use *k* hash functions to set bits in the filter to 1.

Lookup(x):

→ Check whether the bits are 1 for each hash function.

BLOOM FILTERS

Bloom Filter



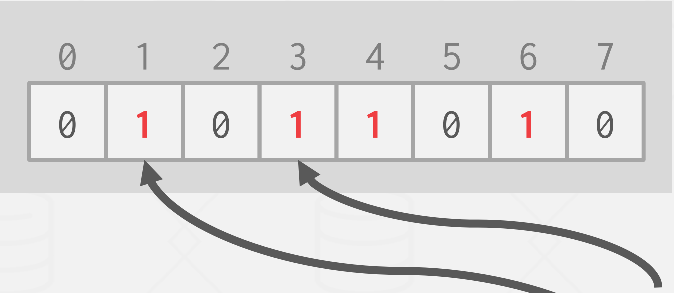
Insert('RZA')

$hash_1$('RZA') = 2222 % 8 = 6

$hash_2$('RZA') = 4444 % 8 = 4

BLOOM FILTERS

Bloom Filter



Insert('RZA')

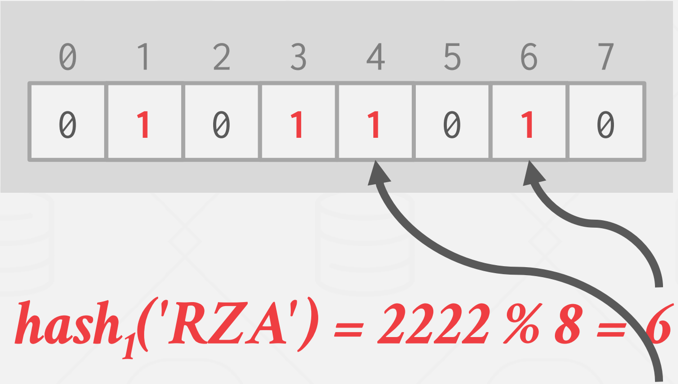
Insert('GZA')

$hash_1('GZA') = 5555 \% 8 = 3$

$hash_2('GZA') = 7777 \% 8 = 1$

BLOOM FILTERS

Bloom Filter



$hash_1$('RZA') = 2222 % 8 = 6

$hash_2$('RZA') = 4444 % 8 = 4

Insert('RZA')

Insert('GZA')

Lookup(RZA') → ***TRUE***

BLOOM FILTERS

Bloom Filter

0	1	2	3	4	5	6	7
0	1	0	1	1	0	1	0

$hash_1$('Raekwon') = 3333 % 8 = 5

$hash_2$('Raekwon') = 8899 % 8 = 3

Insert('RZA')

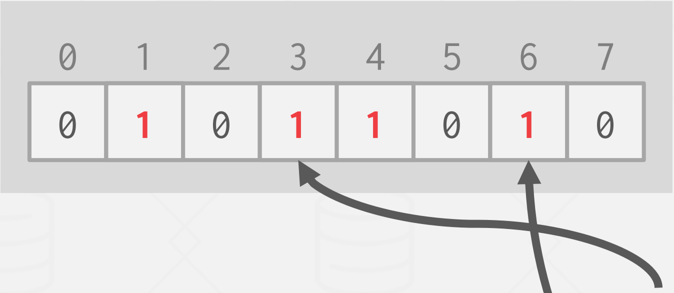
Insert('GZA')

Lookup(RZA') → ***TRUE***

Lookup('Raekwon')→ ***FALSE***

BLOOM FILTERS

Bloom Filter



$hash_1(\text{'ODB'}) = 6699 \,\%\, 8 = 3$

$hash_2(\text{'ODB'}) = 9966 \,\%\, 8 = 6$

Insert('RZA')

Insert('GZA')

Lookup(RZA') → ***TRUE***

Lookup('Raekwon')→ ***FALSE***

Lookup('ODB')→ ***TRUE***

HASH JOIN

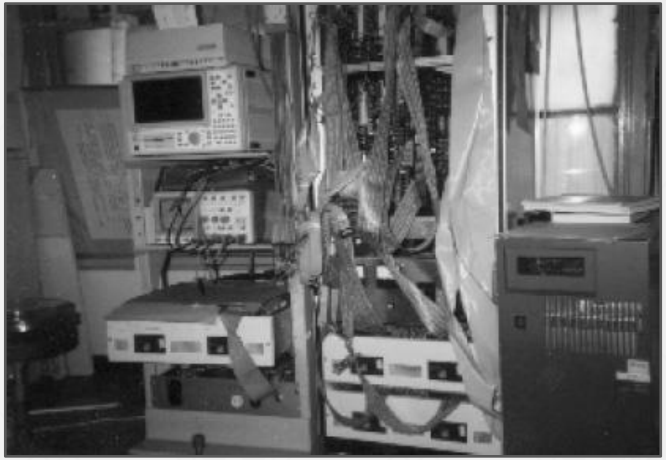
What happens if we do not have enough memory to fit the entire hash table?

We do not want to let the buffer pool manager swap out the hash table pages at random.

PARTITIONED HASH JOIN

Hash join when tables do not fit in memory.
→ **Build Phase:** Hash both tables on the join attribute into partitions.
→ **Probe Phase:** Compares tuples in corresponding partitions for each table.

Sometimes called **GRACE Hash Join**.
→ Named after the GRACE database machine from Japan in the 1980s.



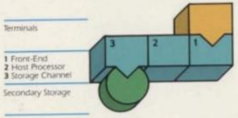
GRACE
University of Tokyo

Britton-Lee's technical achievements have created the Intelligent Data Base Machine, oriented to managers who know the value of a responsive information system. Truly user-oriented—even to people without programming knowledge—the IDM 500 provides some remarkable advantages. Imagine how the features described inside can improve YOUR company's information productivity...

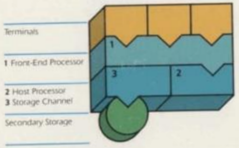
NOW

The IDM 500 A Logical Development

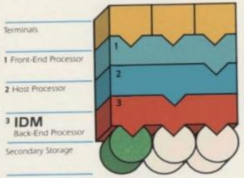
As data systems have evolved, the presence of special-purpose elements has become increasingly important, as these diagrams will illustrate:



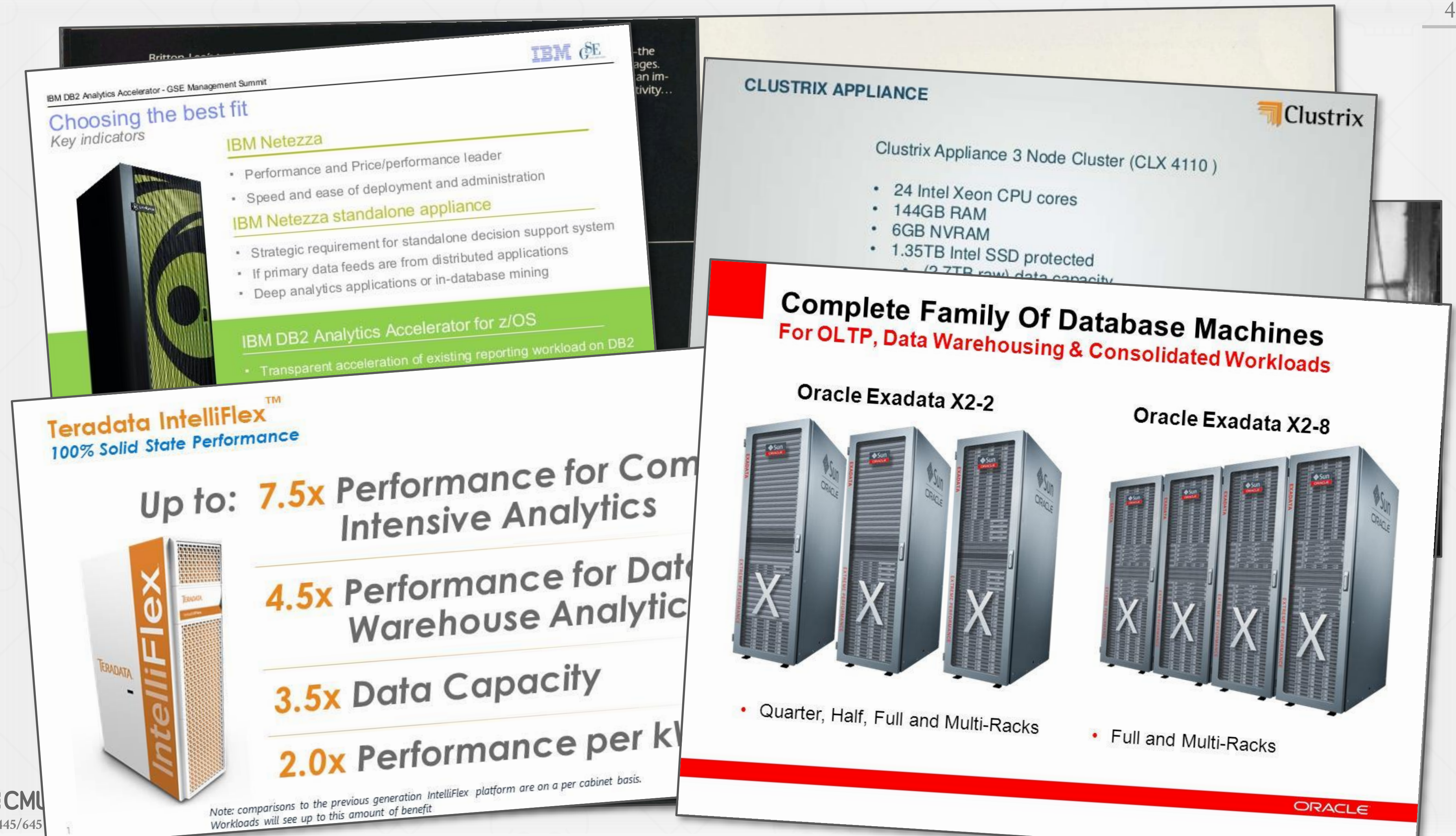
In the 1960's, a single central processing unit (CPU) was required to monitor time-sharing among terminal users; to batch process computing tasks, and to control the access to stored data.

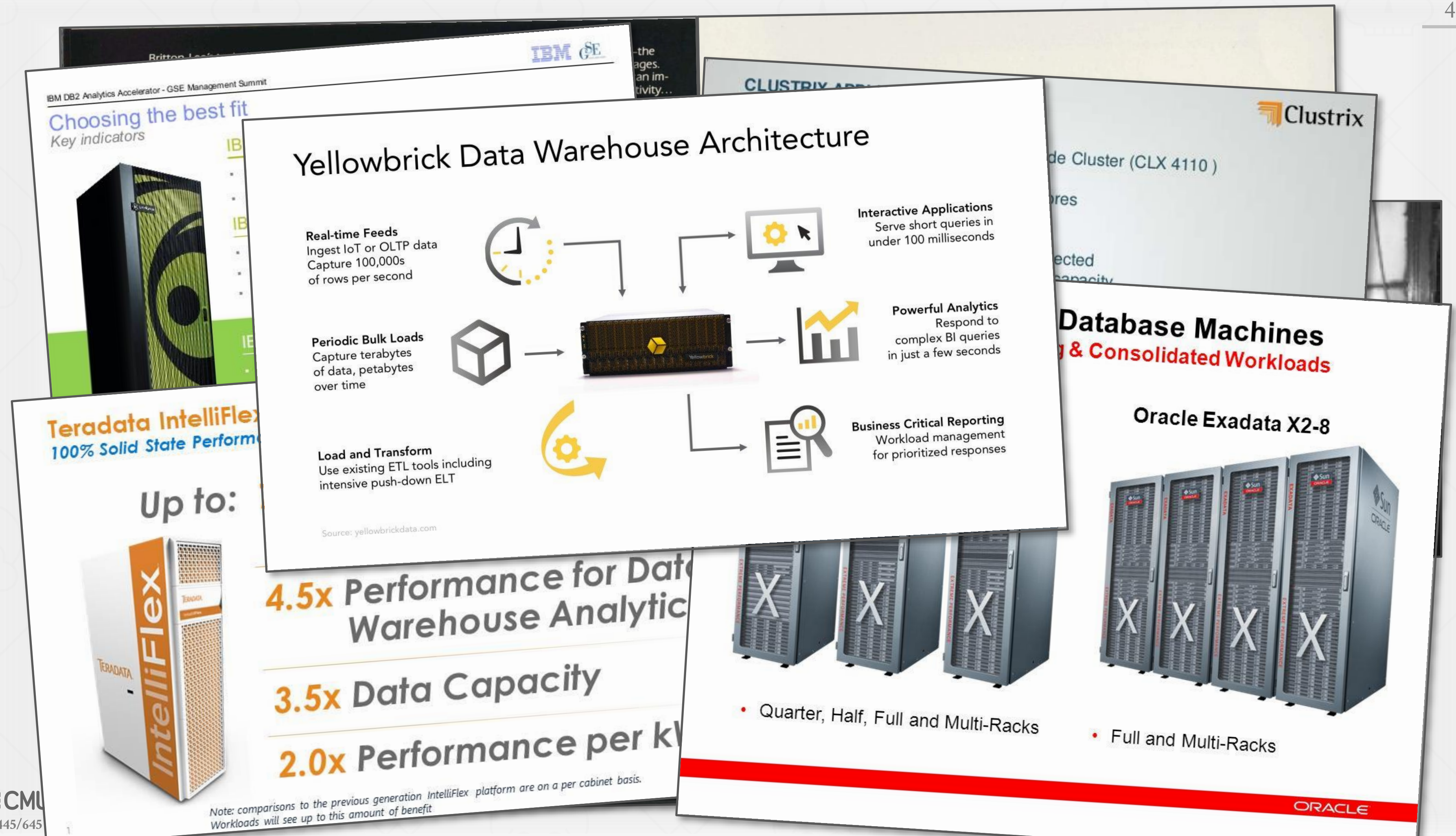


Through the development of front-end communication processors, the work-load on the CPU was reduced. It was then able to perform its basic task of data processing much more efficiently. But the task of managing the data base was still imposed upon it.



Now Britton-Lee's IDM 500 special-purpose, back-end data-base processor brings full efficiency to the host computer and intelligent terminals, so that they can properly perform their correct functions.


IBM DB2 Analytics Accelerator - GSE Management Summit
IBM GSE
Choosing the best fit
Key indicators
IBM Netezza
Performance and Price/performance leader
Speed and ease of deployment and administration
IBM Netezza standalone appliance
Strategic requirement for standalone decision support system
If primary data feeds are from distributed applications
Deep analytics applications or in-database mining
IBM DB2 Analytics Accelerator for z/OS
Transparent acceleration of existing reporting workload on DB2
CLUSTRIX APPLIANCE
Clustrix
Clustrix Appliance 3 Node Cluster (CLX 4110)
24 Intel Xeon CPU cores
144GB RAM
6GB NVRAM
1.35TB Intel SSD protected
Teradata IntelliFlex™
100% Solid State Performance
Up to: 7.5x Performance for Com
Intensive Analytics
4.5x Performance for Dat
Warehouse Analytic
3.5x Data Capacity
2.0x Performance per k
Note: comparisons to the previous generation IntelliFlex platform are on a per cabinet basis.
Workloads will see up to this amount of benefit
Complete Family Of Database Machines
For OLTP, Data Warehousing & Consolidated Workloads
Oracle Exadata X2-2
Oracle Exadata X2-8
Quarter, Half, Full and Multi-Racks
Full and Multi-Racks
ORACLE


IBM GSE
IBM DB2 Analytics Accelerator - GSE Management Summit
Choosing the best fit
Key indicators
Yellowbrick Data Warehouse Architecture
Real-time Feeds
Ingest IoT or OLTP data
Capture 100,000s
of rows per second
Periodic Bulk Loads
Capture terabytes
of data, petabytes
over time
Load and Transform
Use existing ETL tools including
intensive push-down ELT
Interactive Applications
Serve short queries in
under 100 milliseconds
Powerful Analytics
Respond to
complex BI queries
in just a few seconds
Business Critical Reporting
Workload management
for prioritized responses
Source: yellowbrickdata.com
Clustrix
de Cluster (CLX 4110)
Teradata IntelliFlex
100% Solid State Perform
Up to:
4.5x Performance for Data
Warehouse Analytic
3.5x Data Capacity
2.0x Performance per kW
Note: comparisons to the previous generation IntelliFlex platform are on a per cabinet basis.
Workloads will see up to this amount of benefit
IntelliFlex
TERADATA
Database Machines
& Consolidated Workloads
Oracle Exadata X2-8
Quarter, Half, Full and Multi-Racks
Full and Multi-Racks
ORACLE

PARTITIONED HASH JOIN

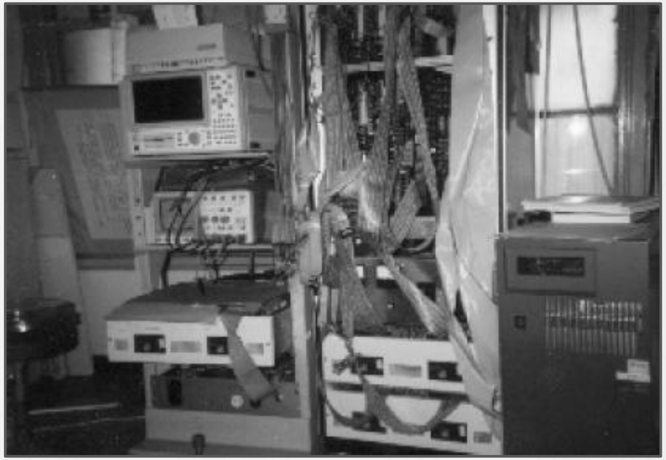
Hash join when tables do not fit in memory.

→ **Build Phase:** Hash both tables on the join attribute into partitions.

→ **Probe Phase:** Compares tuples in corresponding partitions for each table.

Sometimes called **GRACE Hash Join**.

→ Named after the GRACE database machine from Japan in the 1980s.

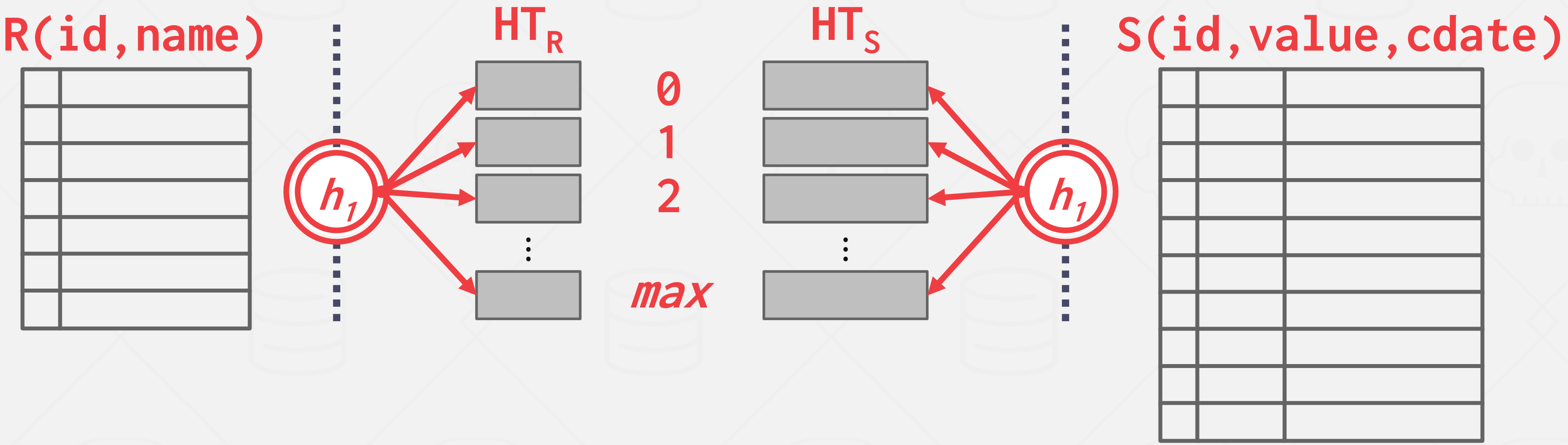


GRACE
University of Tokyo

PARTITIONED HASH JOIN

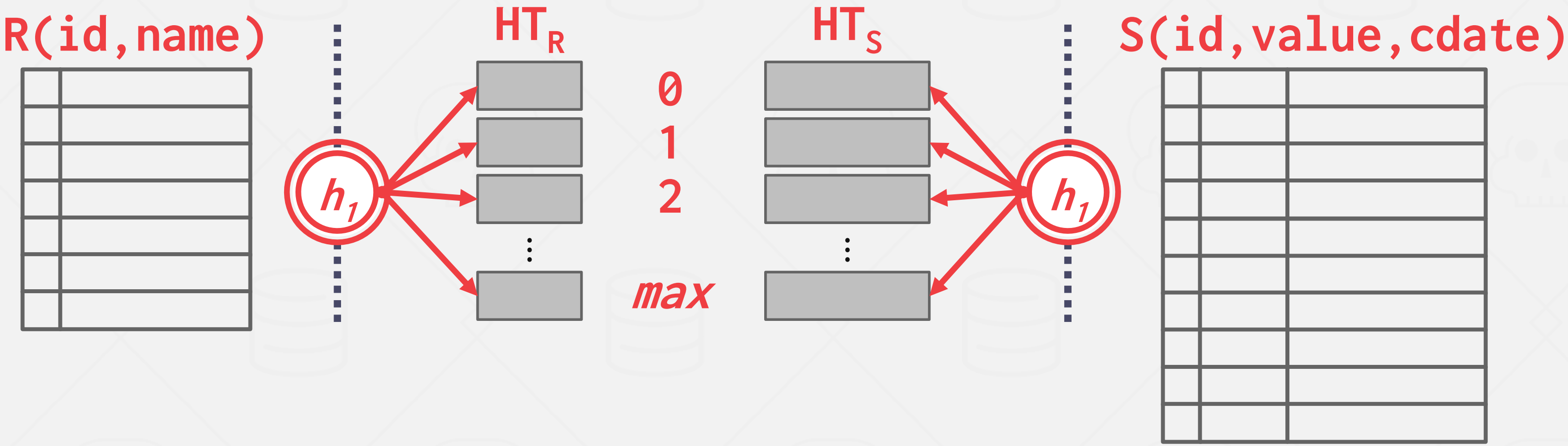
Hash **R** into (0, 1, ..., *max*) buckets.

Hash **S** into the same # of buckets with the same hash function.



PARTITIONED HASH JOIN

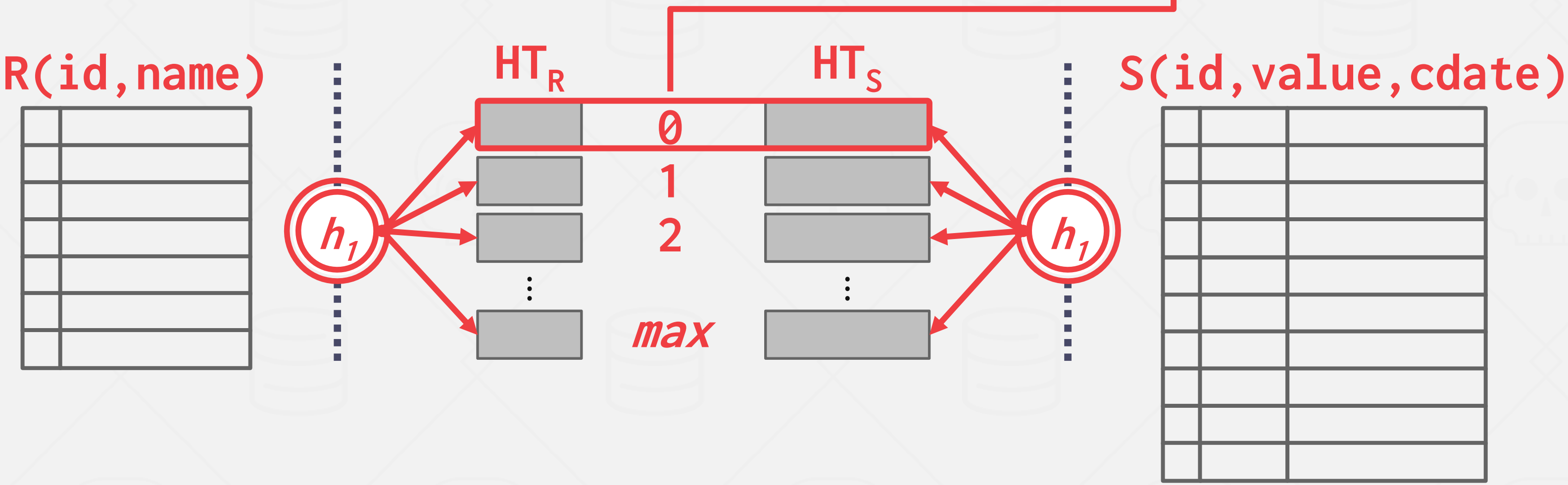
Perform regular hash join on each pair of matching buckets in the same level between **R** and **S**.



PARTITIONED HASH JOIN

Perform regular hash joi matching buckets in the same level between **R** and **S**.

build hash table $HT_{R,0}$ for $bucket_{R,0}$
foreach tuple $s \in bucket_{S,0}$
  output, if $h_2(s) \in HT_{R,0}$



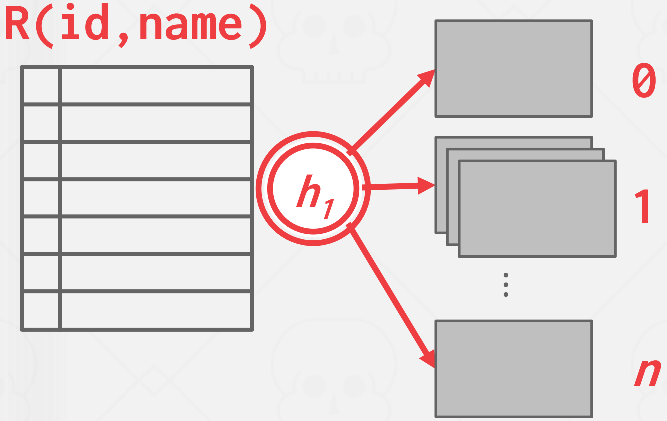
PARTITIONED HASH JOIN

If the buckets do not fit in memory, then use **recursive partitioning** to split the tables into chunks that will fit.

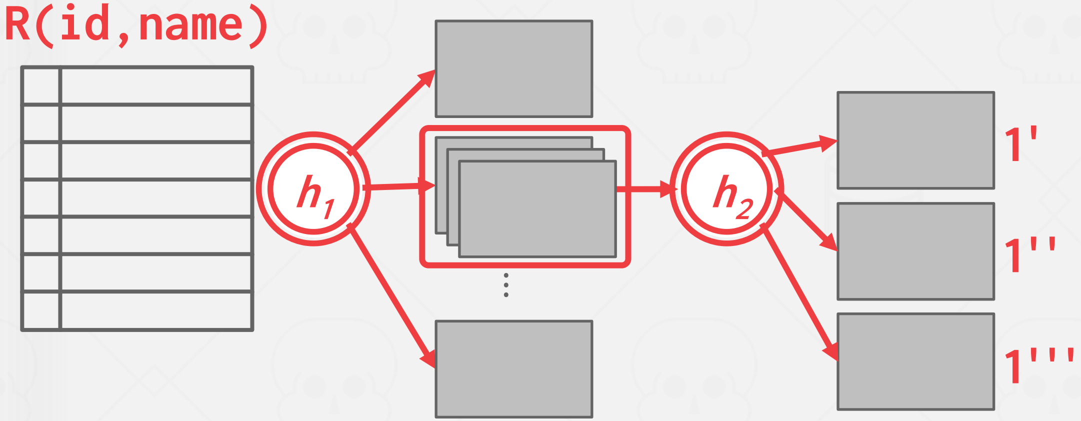
→ Build another hash table for $\mathtt{bucket}_{R,i}$ using hash function h_2 (with $h_2 \neq h_1$).

→ Then probe it for each tuple of the other table's bucket at that level.

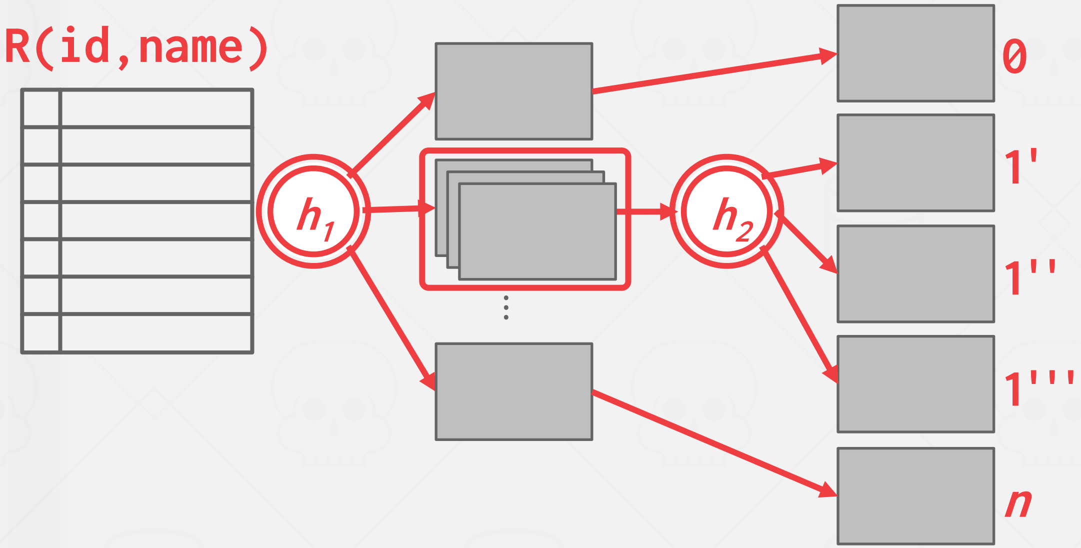
RECURSIVE PARTITIONING



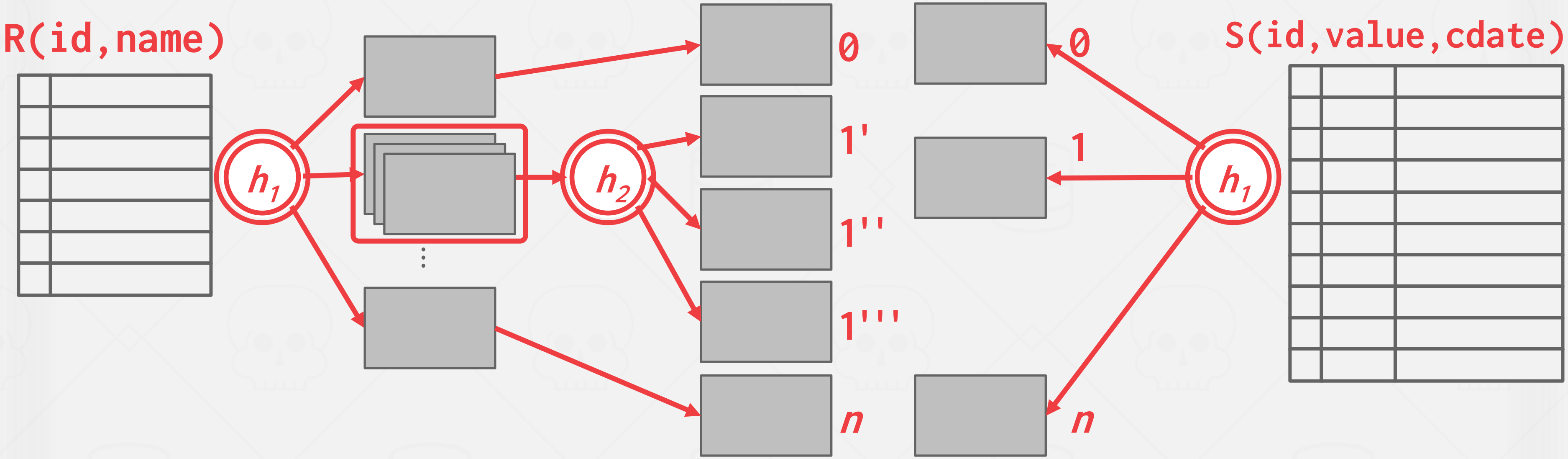
RECURSIVE PARTITIONING



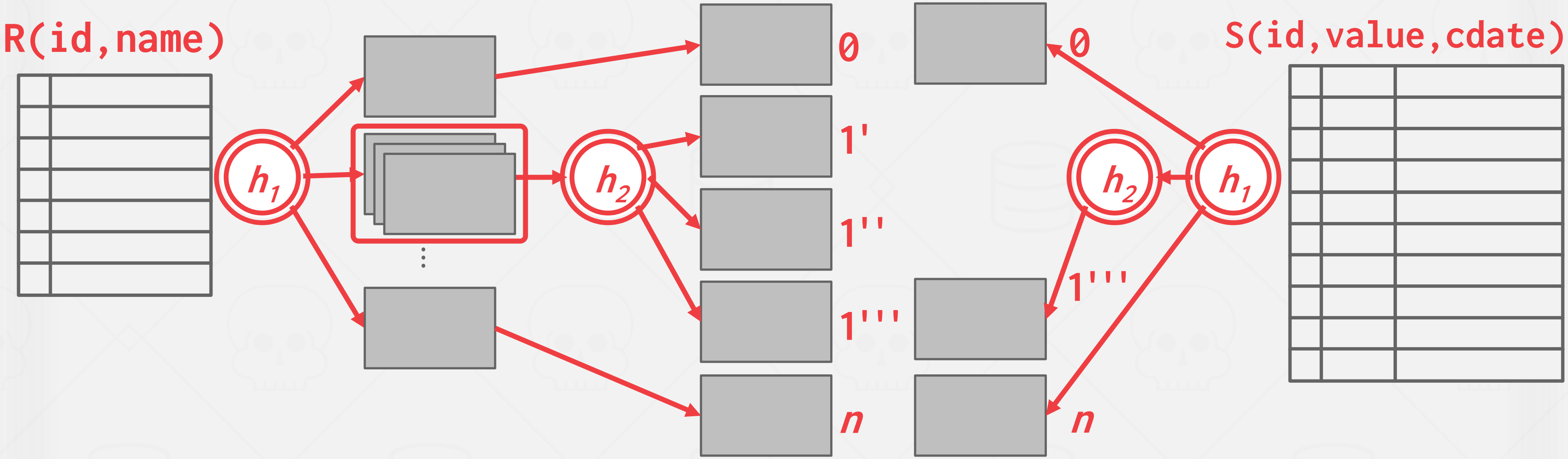
RECURSIVE PARTITIONING



RECURSIVE PARTITIONING



RECURSIVE PARTITIONING



PARTITIONED HASH JOIN

Cost of hash join?
- → Assume that we have enough buffers.
- → Cost: $3(M + N)$

Partitioning Phase:
- → Read+Write both tables
- → $2(M+N)$ IOs

Probing Phase:
- → Read both tables
- → $M+N$ IOs

PARTITIONED HASH JOIN

Example database:

→ M = 1000, m = 100,000

→ N = 500, n = 40,000

Cost Analysis:

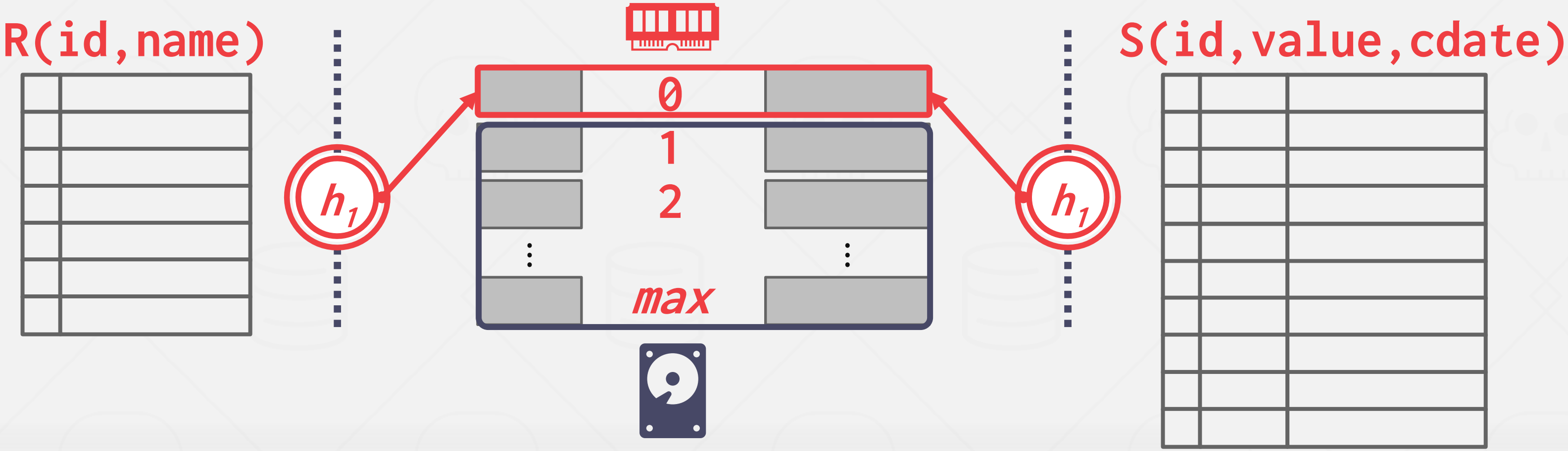
→ $3 \cdot (M + N) = 3 \cdot (1000 + 500) =$ **4,500 IOs**

→ At 0.1 ms/IO, Total time ≈ 0.45 seconds

OPTIMIZATION: HYBRID HASH JOIN

If the keys are skewed, then the DBMS keeps the hot partition in-memory and immediately perform the comparison instead of spilling it to disk.

→ Difficult to get to work correctly. Rarely done in practice.



OBSERVATION

No constraint on the size of inner table.

If the DBMS knows the size of the outer table, then it can use a static hash table.

→ Less computational overhead for build / probe operations.

If we do not know the size, then we must use a dynamic hash table or allow for overflow pages.

JOIN ALGORITHMS: SUMMARY

Algorithm	IO Cost	Example
Simple Nested Loop Join	$M + (m \cdot N)$	1.3 hours
Block Nested Loop Join	$M + (M \cdot N)$	50 seconds
Index Nested Loop Join	$M + (m \cdot C)$	Variable
Sort-Merge Join	$M + N + (\text{sort cost})$	0.75 seconds
Hash Join	$3 \cdot (M + N)$	0.45 seconds

CONCLUSION

Hashing is almost always better than sorting for operator execution.

Caveats:
→ Sorting is better on non-uniform data.
→ Sorting is better when result needs to be sorted.

Good DBMSs use either (or both).

NEXT CLASS

Composing operators together to execute queries.

PROJECT #2

You will build a thread-safe B+tree.

→ Page Layout
→ Data Structure
→ Iterator
→ Latch Crabbing

We define the API for you. You need to provide the method implementations.

https://15445.courses.cs.cmu.edu/fall2022/project2

CHECKPOINT #1

Due Date: October 11th @ 11:59pm
Total Project Grade: 50%

Page Layouts

→ How each node will store its key/values in a page.
→ You only need to support unique keys.

Data Structure (Find + Insert + Delete)

→ Support point queries (single key).
→ Support inserts with node splitting.
→ Support removal of keys with sibling stealing + merging.
→ Does not need to be thread-safe.

CHECKPOINT #2

Due Date: October 23rd @ 11:59pm
Total Project Grade: 50%

Index Iterator

→ Create a STL iterator for range scans.

Concurrent Index

→ Implement latch crabbing/coupling.

DEVELOPMENT HINTS

Follow the textbook semantics and algorithms.

Set the page size to be small (e.g., 512B) when you first start so that you can see more splits/merges.

Make sure that you protect the internal B+Tree `root_page_id` member.

EXTRA CREDIT

Gradescope Leaderboard runs your code with a specialized in-memory version of BusTub.

The top 20 fastest implementations in the class will receive extra credit for this assignment.
→ **#1:** 50% bonus points
→ **#2–10:** 25% bonus points
→ **#11–20:** 10% bonus points

You must pass all the test cases to qualify!

PLAGIARISM WARNING

The homework and projects must be your own original work. They are **not** group assignments.

You may **not** copy source code from other people or the web.

Plagiarism is **not** tolerated. You will get lit up.
→ Please ask me if you are unsure.

See CMU's Policy on Academic Integrity for additional information.

NEXT CLASS

We are finally going to discuss how to execute some queries…