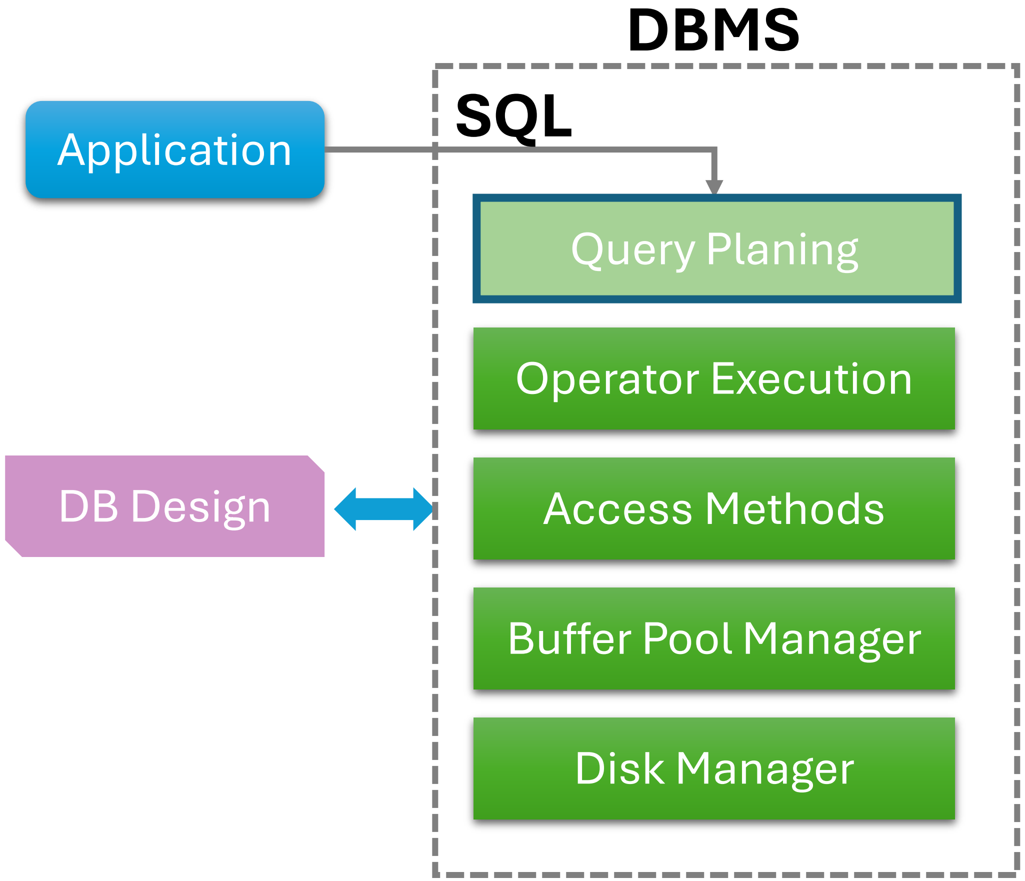
Ack: Prof. Jignesh Patel @ CMU
Prof. Andy Pavlo @CMU

CSC3170
14: Query Planing & Optimization

Chenhao Ma
School of Data Science
The Chinese University of Hong Kong, Shenzhen

This Lecture

- Query Planing & Optimization



SCHOOL OF DATA SCIENCE

Query

```
SELECT distinct ename
FROM Emp E, Dept D
WHERE E.did = D.did AND D.dname = 'Toy'
```

Catalog

clustered unclustered unclustered

EMP (ssn, ename, addr, sal, did)

10,000 records
1,000 pages

clustered unclustered

DEPT (did, dname, floor, mgr)

500 records
50 pages

Total: 2M I/Os

4 reads, 1 write — π_{ename}

2,000 + 4 writes (10K/500 = 20 emps per dept) — $\sigma_{dname = \text{'Toy'}}$

1,000,000 + 2,000 writes (FK join, 10K tuples in temp T2) — $\sigma_{EMP.did = DEPT.did}$

50 + 50,000 + 1,000,000 writes (write to temp file T1) 5 tuples per page in T1 — $\times$

EMP **DEPT**

Query

SELECT distinct ename
FROM Emp E, Dept D
WHERE E.did = D.did AND D.dname = 'Toy'

SCHOOL OF DATA SCIENCE

Total: 54K I/Os

Catalog

clustered unclustered unclustered

EMP (ssn, ename, addr, sal, did)

10,000 records
1,000 pages

clustered unclustered

DEPT (did, dname, floor, mgr)

500 records
50 pages

Read temp T2
4 reads + 1 writes

π_{ename}

Read temp T1
2,000 reads +4 writes

$\sigma_{dname = \text{'Toy'}}$

Page NL, write to temp T1
50 + 50,000 + 2000 writes

$\bowtie_{EMP.did = DEPT.did}$

EMP **DEPT**

Query

SELECT distinct ename
FROM Emp E, Dept D
WHERE E.did = D.did AND D.dname = 'Toy'

w/ Materialization **Total: 7,159 I/Os**

w/ Pipelining **Total: 3,151 I/Os**

Catalog

clustered unclustered unclustered

EMP (ssn, ename, addr, sal, did)

10,000 records
1,000 pages

clustered unclustered

DEPT (did, dname, floor, mgr)

500 records
50 pages

Read temp T2
~~4 reads~~ + 1 writes

π_{ename}

Read temp T1
~~2,000 reads~~ ~~+4 writes~~

$\sigma_{\text{dname = 'Toy'}}$

Sort-merge join (50 buffers)
3*(|Emp|+|Dept|)
= 3150 + ~~2000 writes~~

$\bowtie_{\text{EMP.did = DEPT.did}}$

EMP **DEPT**

Query

SELECT distinct ename
FROM Emp E, Dept D
WHERE E.did = D.did AND D.dname = ‘Toy’

Total: 37 I/Os

Catalog

clustered | unclustered | unclustered

EMP (ssn, ename, addr, sal, did)

10,000 records
1,000 pages

clustered | unclustered

DEPT (did, dname, floor, mgr)

500 records
50 pages

Read temp T2
4 reads + 1 writes

π_{ename}

Read temp T1, **NL-IDX Join**
1 + 3 (idx) + 20 (ptr chase) + 4 writes

$\sigma_{dname = ‘Toy’}$

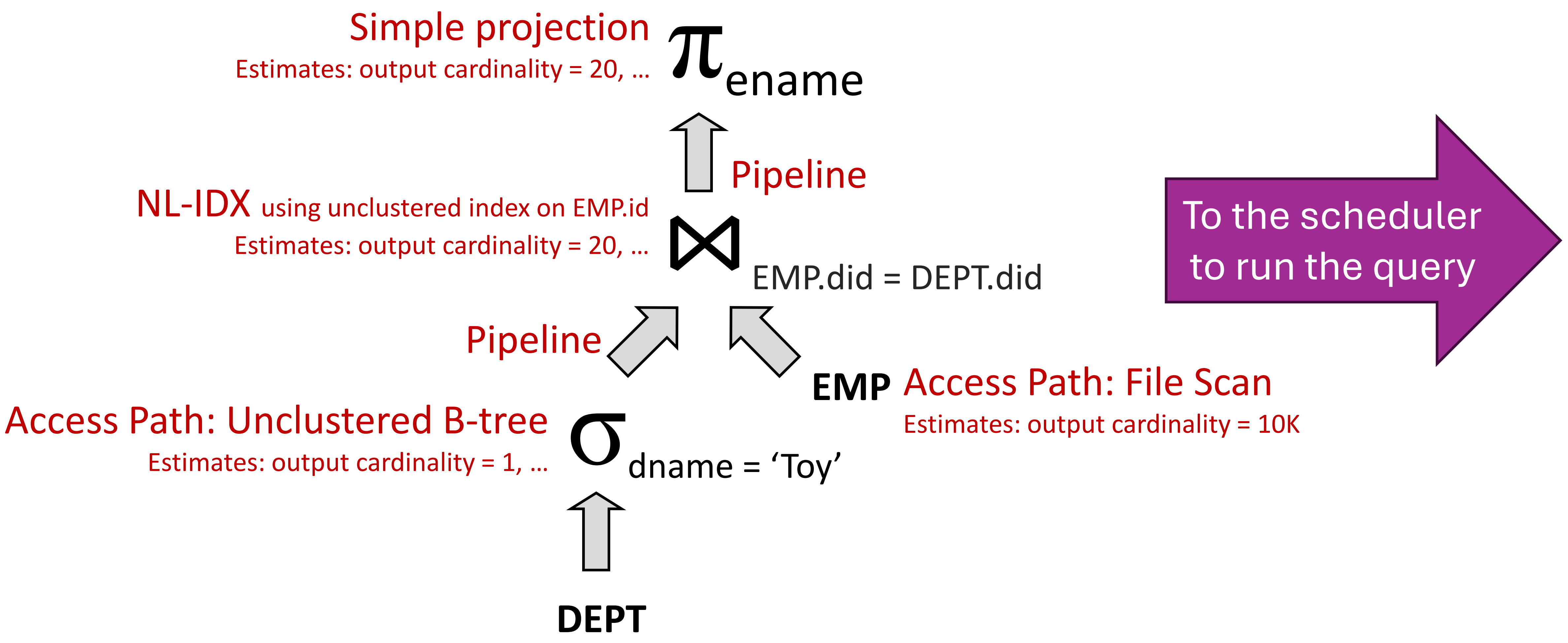
$\bowtie$

Access: **Index (name)**
3 reads + 1 write

DEPT EMP | DEPT EMP

Swap

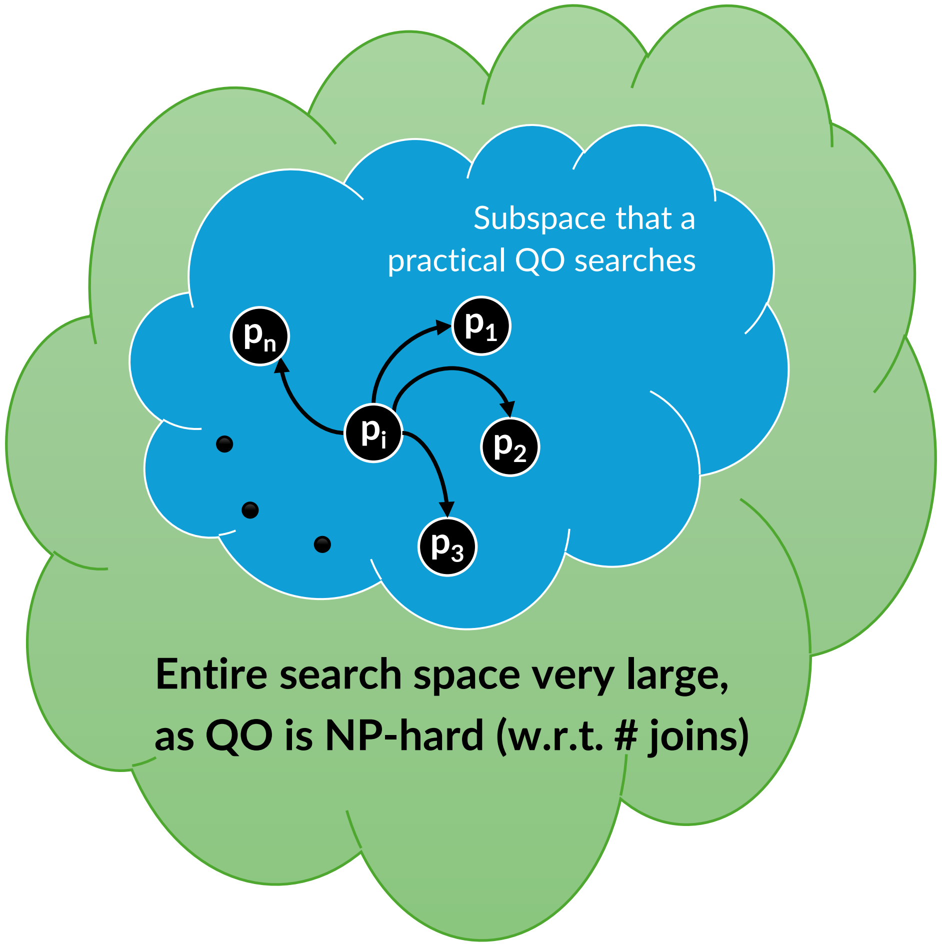
Annotated RA Tree a.k.a. The Physical Plan



Query Optimization (QO)

1. Identify candidate <u>equivalent</u> trees (logical). It is an NP-hard problem, so the space is large.
2. For each candidate, find the execution plan tree (physical). We need to **estimate** the cost of each plan.
3. Choose the best overall (physical) plan.

- **Practically: Choose from a subset of all possible plans.**



Logical VS. Physical Plans

- The optimizer generates a mapping of a **<u>logical</u>** algebra expression to the optimal equivalent physical algebra expression.

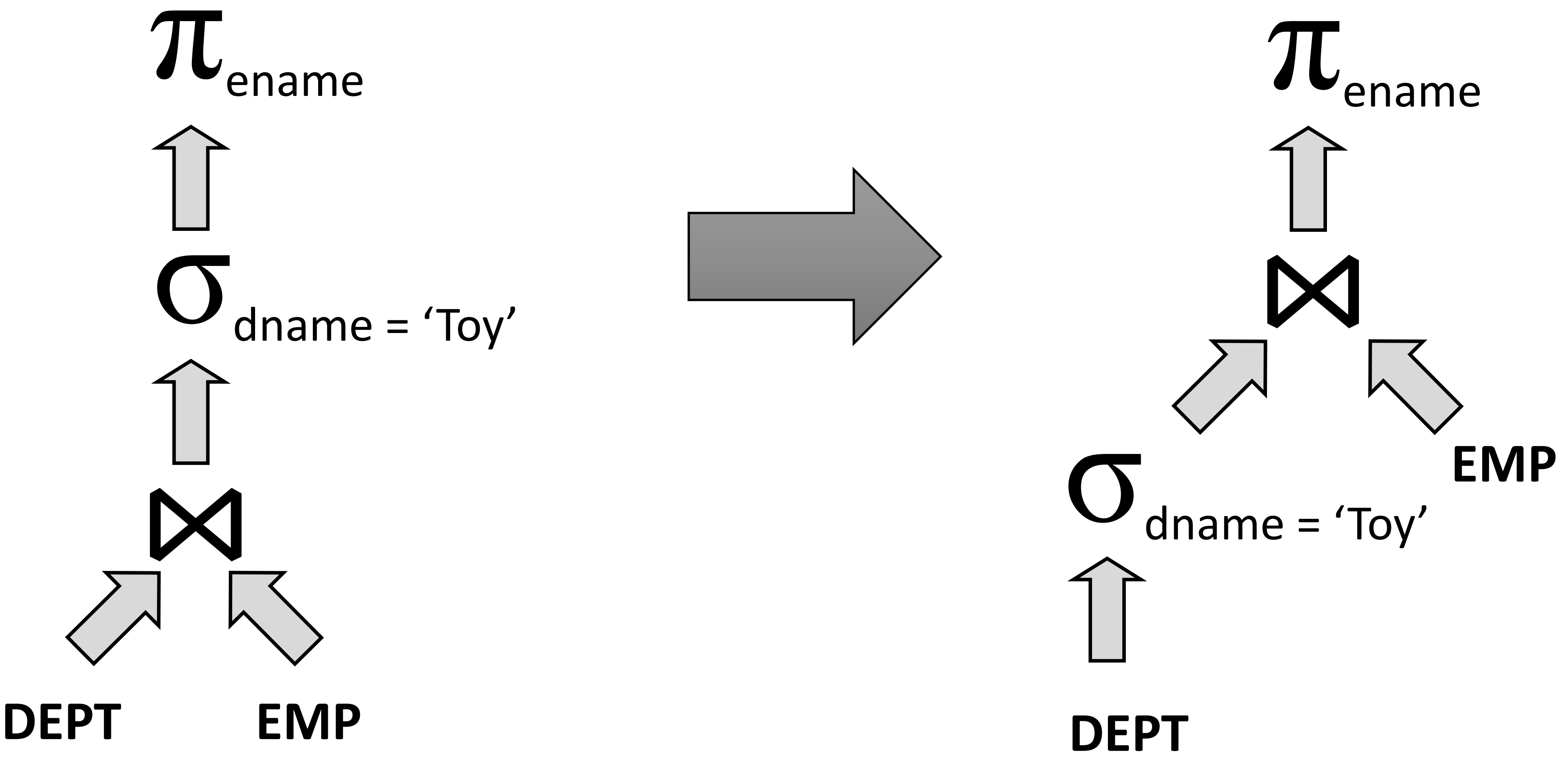
- **<u>Physical</u>** operators define a specific execution strategy using an access path.
  - They can depend on the physical format of the data that they process (i.e., sorting, compression).
  - Not always a 1:1 mapping from logical to physical.

Query Optimization

- **Heuristics / Rules**
  - Rewrite the query to remove (guessed) inefficiencies; e.g, always do selections first, or push down projections as early as possible.
  - These techniques may need to examine catalog, but they do not need to examine data.

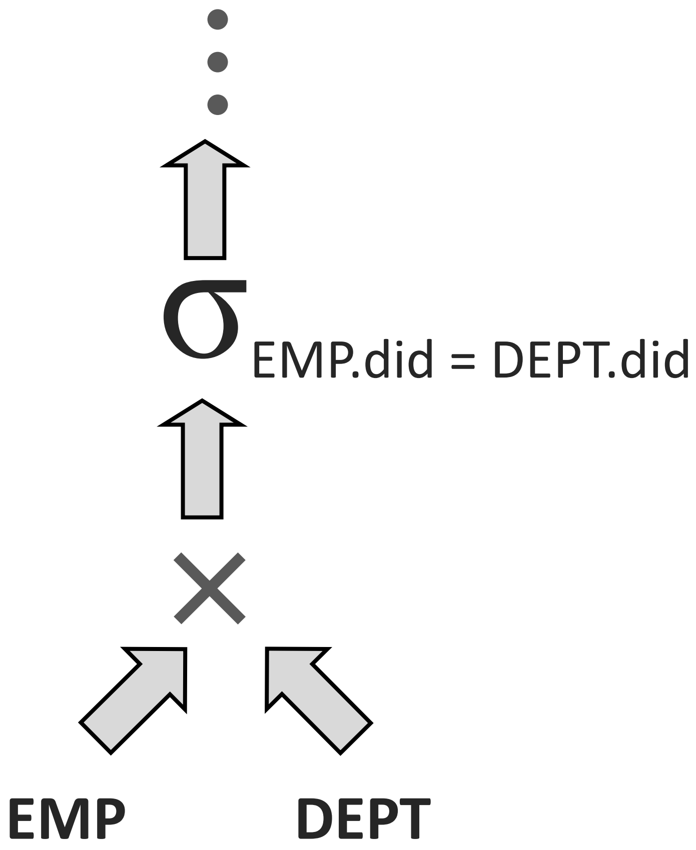
- **Cost-based Search**
  - Use a model to estimate the cost of executing a plan.
  - Enumerate multiple equivalent plans for a query and pick the one with the lowest cost.

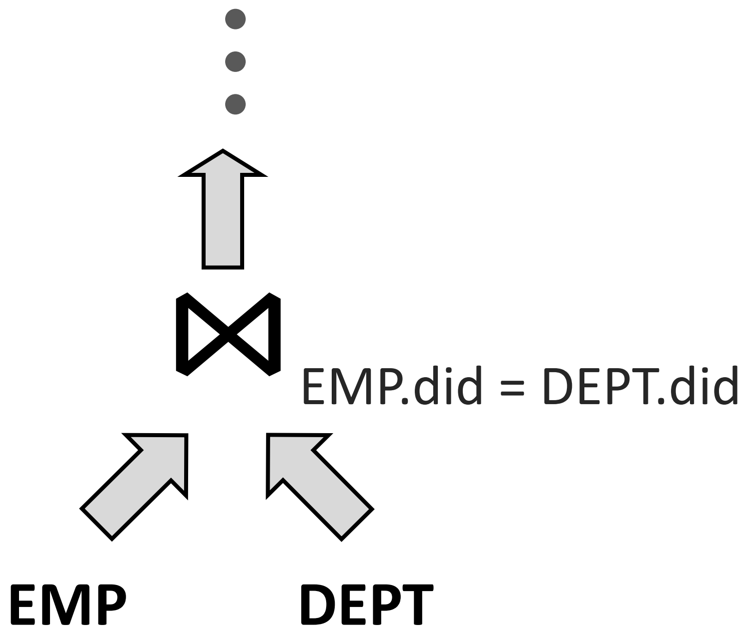
Predicate Pushdown



$$\pi_{ename}\left(\sigma_{dname\,=\,\text{'Toy'}}\left(DEPT \bowtie EMP\right)\right) \xrightarrow{\text{rewrite}} \pi_{ename}\left(EMP \bowtie \sigma_{dname\,=\,\text{'Toy'}}\left(DEPT\right)\right)$$

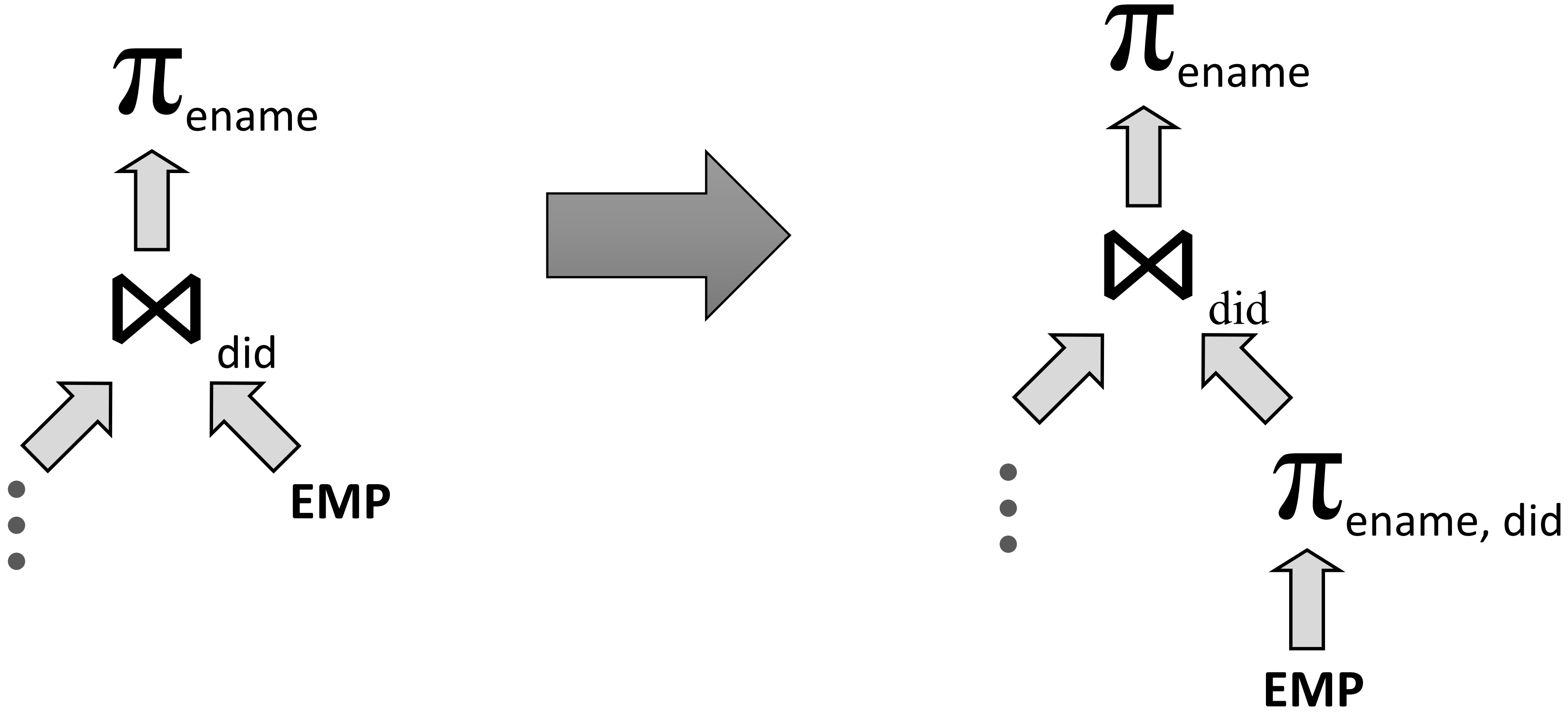
Replace Cartesian Product





$\ldots(\sigma_{\text{DEPT.did = EMP.did}}(\text{DEPT X EMP}))$ rewrite $\ldots(\text{EMP} \bowtie_{\text{did}} \text{DEPT})$

Projection Pushdown

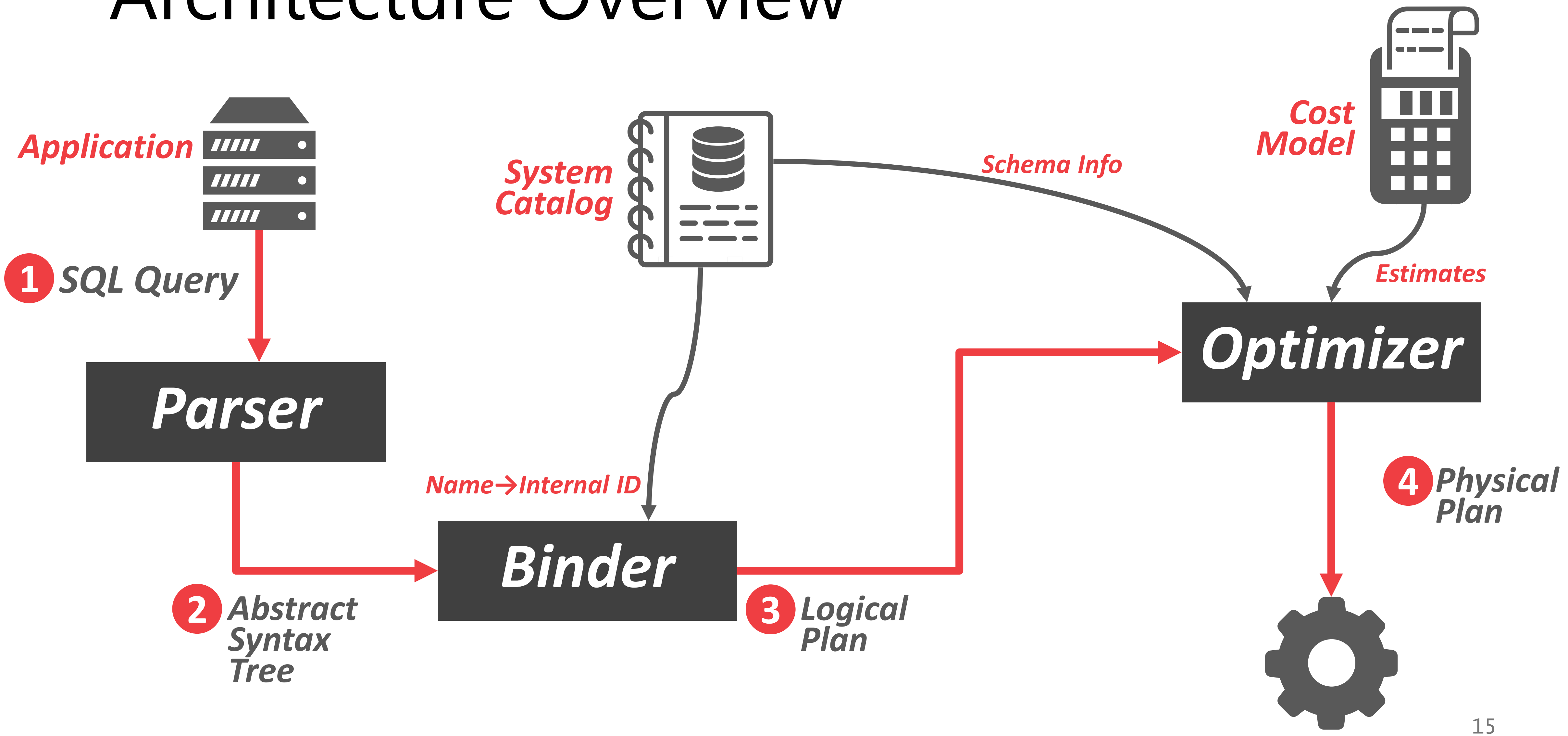


$\pi_{\text{EMP.ename}}(\ldots \bowtie_{\text{did}} \text{EMP})$ rewrite $\pi_{\text{EMP.ename}}(\ldots \bowtie_{\text{did}} (\pi_{\text{ename, did}} \text{EMP}))$

Equivalence

- $\sigma_{P1}(\sigma_{P2}(R)) \equiv \sigma_{P2}(\sigma_{P1}(R))$ (σ commutativity)
- $\sigma_{P1 \wedge P2 \ldots \wedge Pn}(R) \equiv \sigma_{P1}(\sigma_{P2}(\ldots \sigma_{Pn}(R)))$ (cascading σ)
- $\prod_{a1}(R) \equiv \prod_{a1}(\prod_{a2}(\ldots\prod_{ak}(R)\ldots))$, $a_i \subseteq a_{i+1}$ (cascading $\prod$)
- $R \bowtie S \equiv S \bowtie R$ (join commutativity)
- $R \bowtie (S \bowtie T) \equiv (R \bowtie S) \bowtie T$ (join associativity)
- $\sigma_P(R \times S) \equiv (R \bowtie_P S)$, if P is a join predicate
- $\sigma_P(R \times S) \equiv \sigma_{P1}(\sigma_{P2}(R) \bowtie_{P4} \sigma_{P3}(S))$, where $P = p1 \wedge p2 \wedge p3 \wedge p4$
- $\prod_{A1,A2,\ldots An}(\sigma_P(R)) \equiv \prod_{A1,A2,\ldots An}(\sigma_P(\prod_{A1,\ldots An,\, B1,\ldots BM}R))$, where B1 … BM are columns in P
- …

Architecture Overview



Query Optimization

Examples: predicate pushdown, replace cartesian product, projection pushdown …

- **Heuristics / Rules**
  - Rewrite the query to remove (guessed) inefficiencies; e.g, always do selections first, or push down projections as early as possible.
  - These techniques may need to examine catalog, but they do not need to examine data.

- **Cost-based Search**
  - Use a model to estimate the cost of executing a plan.
  - Enumerate multiple equivalent plans for a query and pick the one with the lowest cost.

Cost-based Search

Cost-based Query Optimization

- Let's start with a certain style of QO: cost-based, bottom-up QO (the classic System-R optimizer approach)
- Approach: Enumerate different plans for the query and estimate their costs.
  - Single relation.
  - Multiple relations.
  - Nested sub-queries.
- It chooses the best plan it has seen for the query after exhausting all plans or some timeout.

Single-relation Query Planing

- Pick the best access method.
  - Sequential Scan
  - Binary Search (clustered indexes)
  - Index Scan
- Predicate evaluation ordering.

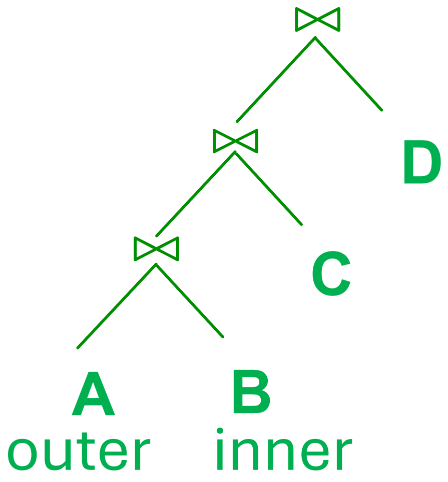
- Simple heuristics are often good enough for this.

System R Optimizer

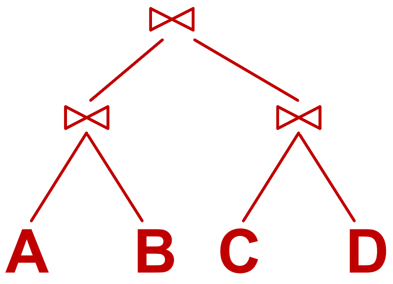
- Break the query into blocks and generate the logical operators for each block.
- For each logical operator, generate a set of physical operators that implement it.
  - All combinations of join algorithms and access paths
- Then, iteratively construct a “left-deep” join tree that minimizes the estimated amount of work to execute the plan.


Selinger

A left-deep tree



A bushy tree



System-R optimizer does NOT consider this “shape”

SCHOOL OF DATA SCIENCE

System R Optimizer

```
SELECT ARTIST.NAME
  FROM ARTIST, APPEARS, ALBUM
 WHERE ARTIST.ID=APPEARS.ARTIST_ID
   AND APPEARS.ALBUM_ID=ALBUM.ID
   AND ALBUM.GENRE="Blues"
 ORDER BY ARTIST.ID
```

ARTIST: Sequential Scan
APPEARS: Sequential Scan
ALBUM: Index Look-up on **GENRE**

Step #1: Choose the best access paths to each table

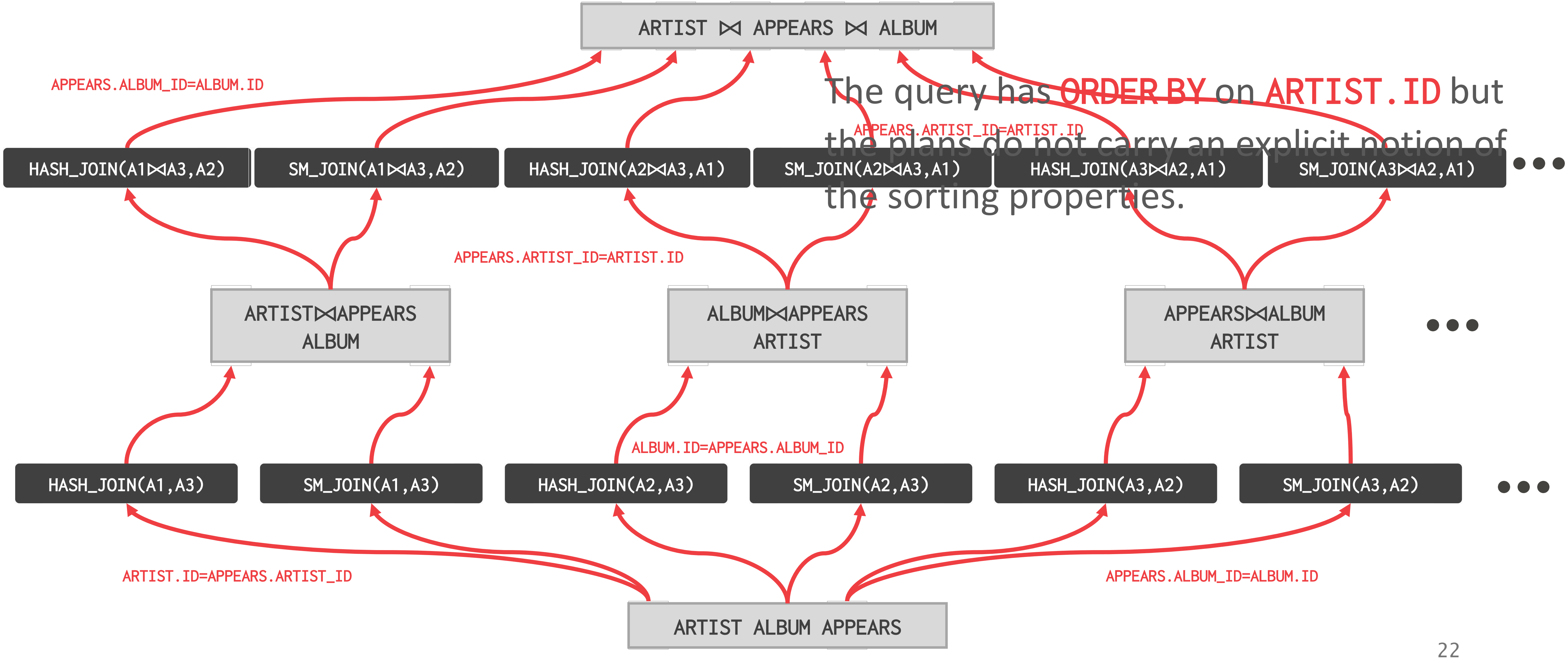
Step #2: Enumerate all possible join orderings for tables

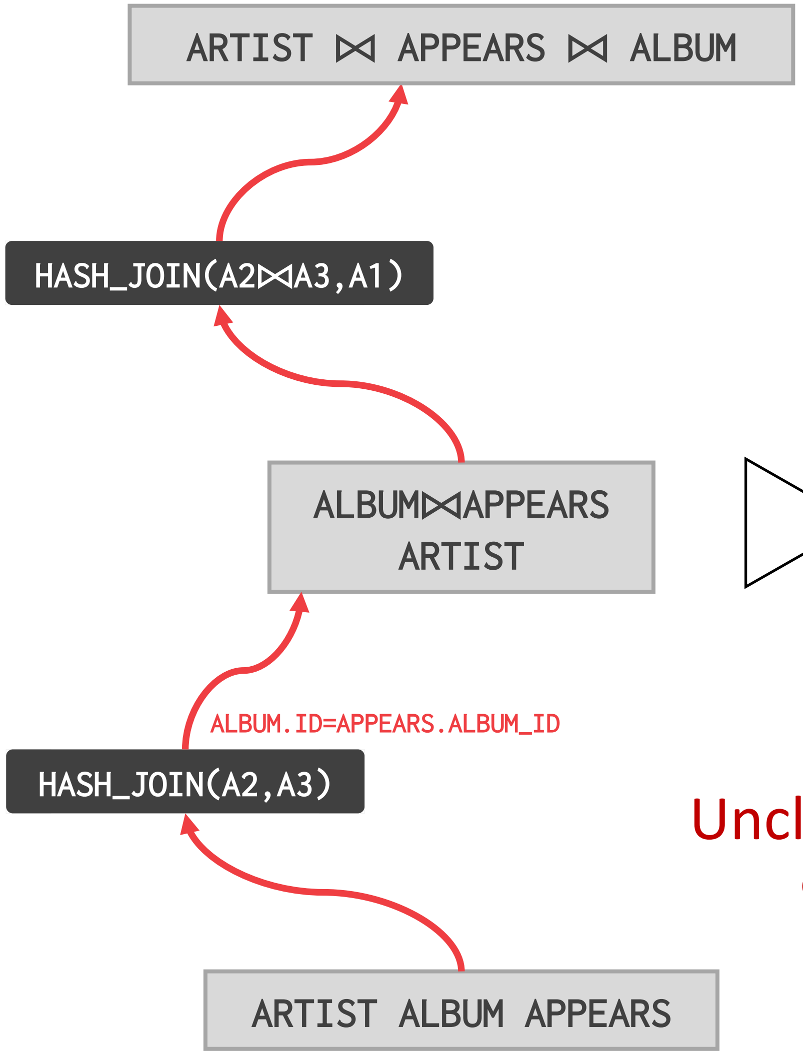
Step #3: Determine the join ordering with the lowest cost

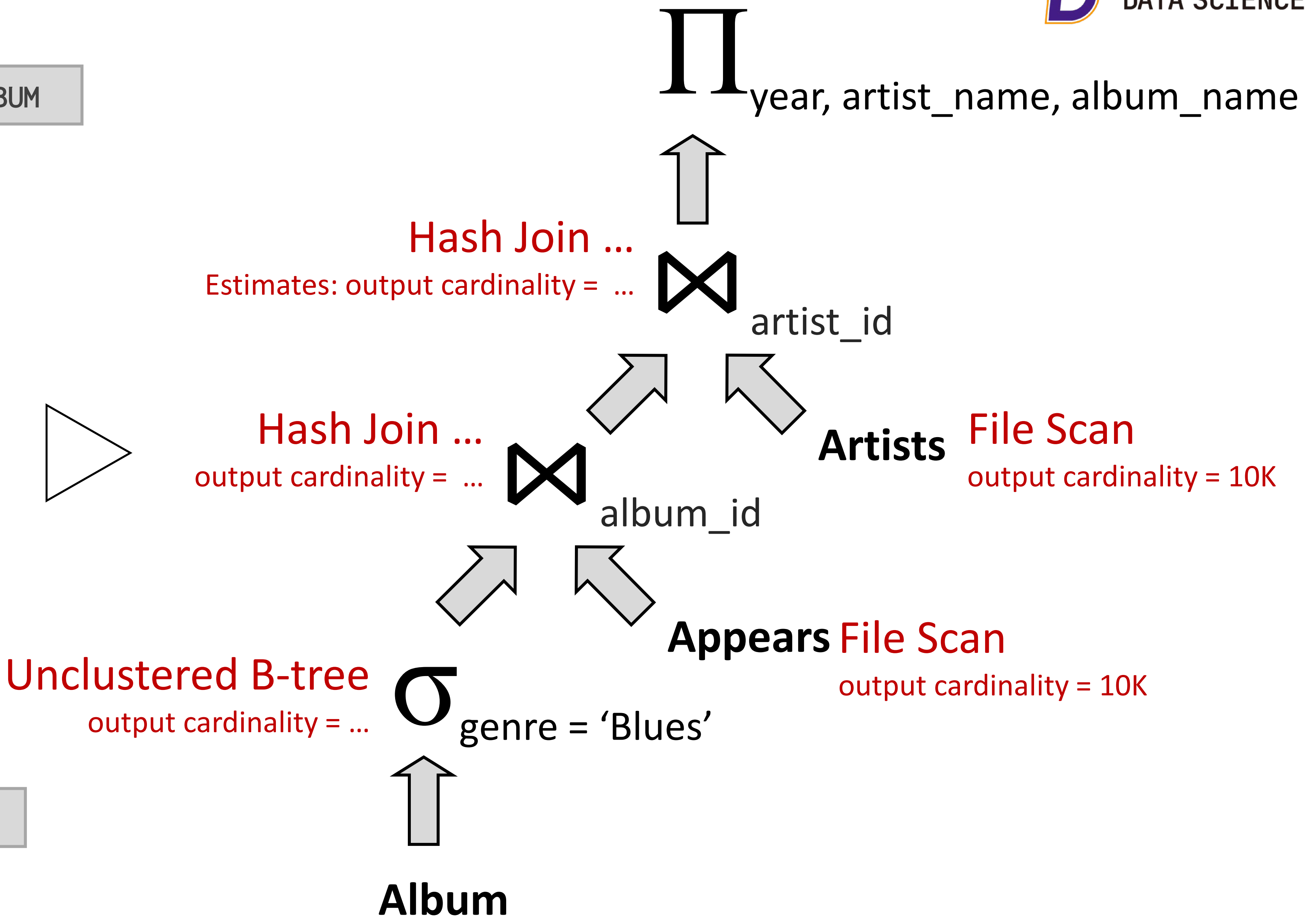
ARTIST	⋈	APPEARS	⋈	ALBUM
APPEARS	⋈	ALBUM	⋈	ARTIST
ALBUM	⋈	APPEARS	⋈	ARTIST
APPEARS	⋈	ARTIST	⋈	ALBUM
ARTIST	×	ALBUM	⋈	APPEARS
ALBUM	×	ARTIST	⋈	APPEARS
⋮		⋮		⋮

System R Optimizer

The query has **ORDER BY** on **ARTIST.ID** but the plans do not carry an explicit notion of the sorting properties.




ARTIST ⋈ APPEARS ⋈ ALBUM
HASH_JOIN(A2⋈A3,A1)
ALBUM⋈APPEARS
ARTIST
ALBUM.ID=APPEARS.ALBUM_ID
HASH_JOIN(A2,A3)
ARTIST ALBUM APPEARS


Π year, artist_name, album_name
Hash Join …
Estimates: output cardinality = …
⋈ artist_id
Hash Join …
output cardinality = …
⋈ album_id
Artists
File Scan
output cardinality = 10K
Appears
File Scan
output cardinality = 10K
Unclustered B-tree
output cardinality = …
σ genre = 'Blues'
Album

Multi-Relation Query Planning

- **Choice #1: Bottom-up Optimization**
  - Start with nothing and then build up the plan to get to the outcome that you want.

We just saw an example of this, the System R approach

- **Choice #2: Top-down Optimization**
  - Start with the outcome that you want, and then work down the tree to find the optimal plan that gets you to that goal.

Bottom-up Optimization

- Use static rules to perform initial optimization.
  Then use dynamic programming to determine
  the best join order for tables using a divide-and-conquer search method

- **Examples**: IBM System R, DB2, MySQL, Postgres, most open-source DBMSs.

Top-down Optimization

- Start with a logical plan of what we want the query to be. Perform a branch-and-bound search to traverse the plan tree by converting logical operators into physical operators.
  - Keep track of global best plan during search.
  - Treat physical properties of data as first-class entities during planning.

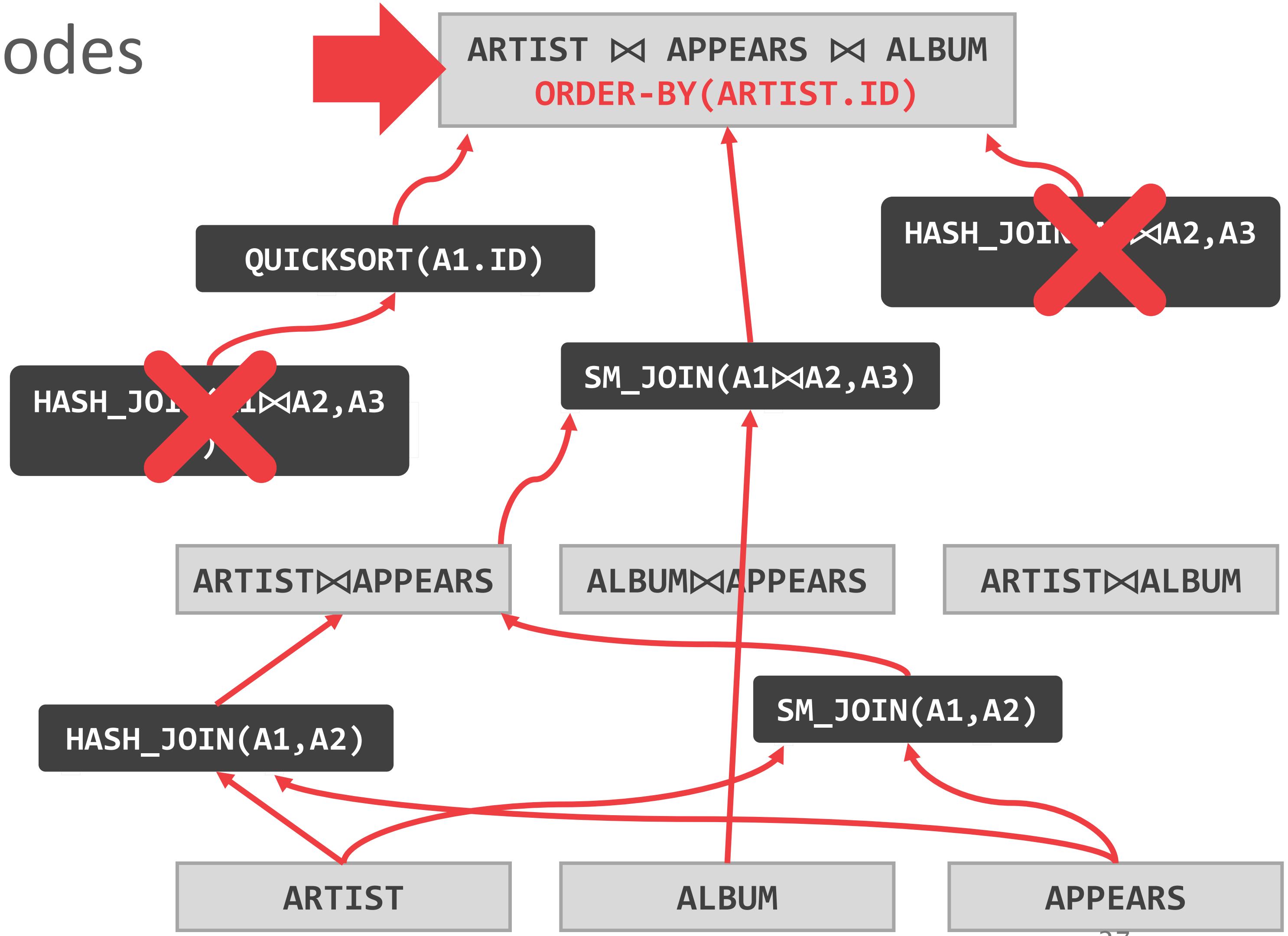
- **Example**: MSSQL, Greenplum, CockroachDB

Top-down Optimization

Invoke rules to create new nodes and traverse the tree.

- →**Logical→Logical:**
  JOIN(A,B) to JOIN(B,A)
- →**Logical→Physical:**
  JOIN(A,B) to HASH_JOIN(A,B)

Can create "enforcer" rules that require input to have certain properties.



Life so far … single block QO

- Often, we get nested queries.
  - We could optimize each block using the methods we have discussed.
  - However, this may be inefficient since we optimize each block separately without a global approach.
- What if we could flatten a nested query into a single block and optimize it?
  - Then, apply single-block query optimization methods.
  - Even if one can't flatten to a single block, flattening to <u>fewer</u> blocks is still beneficial.

Nested Quries

Nested Quries

- The DBMS treats nested sub-queries in the where clause as functions that take parameters and return a single value or set of values.
- Two Approaches:
  - Rewrite to de-correlate and/or flatten them.
  - Decompose nested query and store results in a temporary table.

Nested Sub-queries: Rewrite

```
SELECT name FROM sailors AS S
 WHERE EXISTS (
    SELECT * FROM reserves AS R
     WHERE S.sid = R.sid
       AND R.day = '2022-10-25'
 )
```

```
SELECT name
  FROM sailors AS S, reserves AS R
 WHERE S.sid = R.sid
   AND R.day = '2022-10-25'
```

Decomposing Queries

- For harder queries, the optimizer breaks up queries into blocks and then concentrates on one block at a time.
- Sub-queries are written to temporary tables that are discarded after the query finishes.

Decomposing Queries

Inner Block

```
SELECT MAX(rating) FROM sailors
```

```
SELECT S.sid, MIN(R.day)
  FROM sailors S, reserves R, boats B
 WHERE S.sid = R.sid
   AND R.bid = B.bid
   AND B.color = 'red'
   AND S.rating = ###

 GROUP BY S.sid
HAVING COUNT(*) > 1
```

Outer Block

Nested Block

Expression Rewriting

Expression Rewriting

- An optimizer transforms a query's expressions (e.g., WHERE/ON clause predicates) into the minimal set of expressions.

- Implemented using if/then/else clauses or a pattern-matching rule engine.
  - Search for expressions that match a pattern.
  - When a match is found, rewrite the expression.
  - Halt if there are no more rules that match.

Expression Rewriting

- Impossible / Unnecessary Predicates

```
SELECT * FROM A WHERE false;
```

```
SELECT * FROM A WHERE false;
```

```
SELECT * FROM A WHERE RANDOM() IS NULL;
```

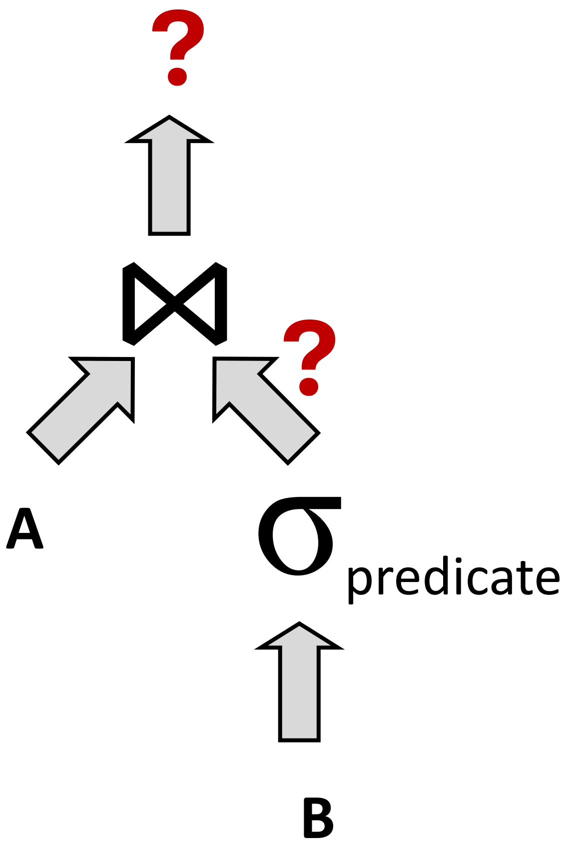
- Merging Predicates

```
SELECT * FROM A
 WHERE val BETWEEN 1 AND 150;
    OR val BETWEEN 50 AND 150;
```

Cost Estimation

How do we calculate the cost of the plans?

- We have formulas for the operator algorithms (e.g. the cost formulaes for hash join, sort merge join, ...), but we also need to estimate the size of the output that an operator produces.



Cost Estimation

- The DBMS uses a cost model to predict the behavior of a query plan given a database state.
  - This is an internal cost that allows the DBMS to compare one plan with another.
- It is too expensive to run every possible plan to determine this information, so the DBMS needs a way to derive this information.

Cost Model Components

- **Choice #1: Physical Costs**
  - Predict CPU cycles, I/O, cache misses, RAM consumption, network messages…
  - Depends heavily on hardware.
- **Choice #2: Logical Costs**
  - Estimate output size per operator.
  - Independent of the operator algorithm.
  - Need estimations for operator result sizes.

SCHOOL OF DATA SCIENCE

Postgres Cost Model

- Uses a combination of CPU and I/O costs that are weighted by "magic" constant factors.

- Default settings are obviously for a disk-resident database without a lot of memory:
  - Processing a tuple in memory is **400x** faster than reading a tuple from disk.
  - Sequential I/O is **4x** faster than random I/O.

19.7.2. Planner Cost Constants

The *cost* variables described in this section are measured on an arbitrary scale. Only their relative values matter, hence scaling them all up or down by the same factor will result in no change in the planner's choices. By default, these cost variables are based on the cost of sequential page fetches; that is, `seq_page_cost` is conventionally set to `1.0` and the other cost variables are set with reference to that. But you can use a different scale if you prefer, such as actual execution times in milliseconds on a particular machine.

> **Note:** Unfortunately, there is no well-defined method for determining ideal values for the cost variables. They are best treated as averages over the entire mix of queries that a particular installation will receive. This means that changing them on the basis of just a few experiments is very risky.

`seq_page_cost` (`floating point`)

Sets the planner's estimate of the cost of a disk page fetch that is part of a series of sequential fetches. The default is 1.0. This value can be overridden for tables and indexes in a particular tablespace by setting the tablespace parameter of the same name (see ALTER TABLESPACE).

`random_page_cost` (`floating point`)

Statistics

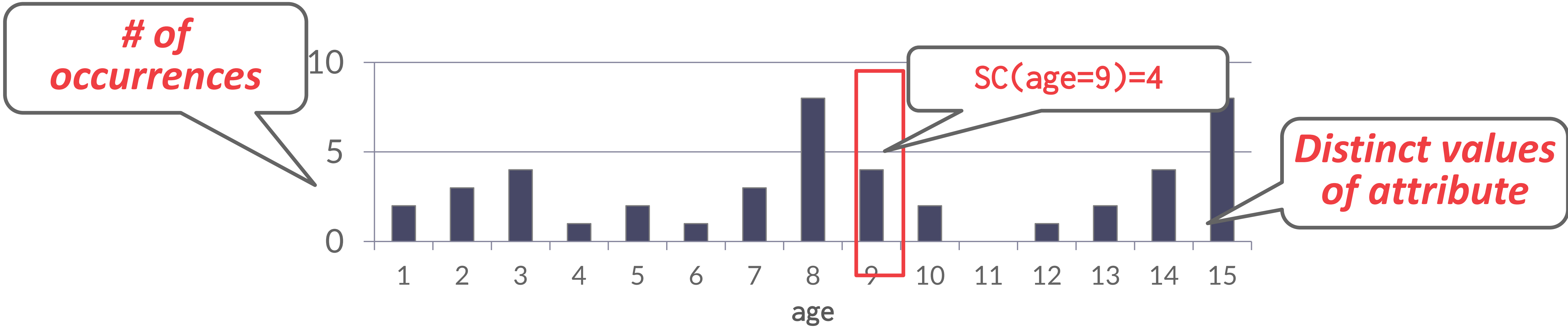
- The DBMS stores internal statistics about tables, attributes, and indexes in its internal catalog.
- Different systems update them at different times.

- Manual invocations:
  - Postgres/SQLite: `ANALYZE`
  - Oracle/MySQL: `ANALYZE TABLE`
  - SQL Server: `UPDATE STATISTICS`
  - DB2: `RUNSTATS`

Selection Cardinality

- The **selectivity** (sel) of a predicate P is the fraction of tuples that qualify.
- **Equality Predicate:** A=constant
  - sel(A=constant) = #occurences/|R|
  - Example: sel(age=9) = 4/45

```
SELECT * FROM people
 WHERE age = 9
```



Selection Cardinality

- **Assumption #1: Uniform Data**
  - The distribution of values (except for the heavy hitters) is the same.
- **Assumption #2: Independent Predicates**
  - The predicates on attributes are independent
- **Assumption #3: Inclusion Principle**
  - The domain of join keys overlap such that each key in the inner relation will also exist in the outer table.

Correlated Attributes

- Consider a database of automobiles:
  - # of Makes = 10, # of Models = 100
- And the following query:
  - `(make="Honda" AND model="Accord")`
- With the independence and uniformity assumptions, the selectivity is:
  - $1/10 \times 1/100 = 0.001$
- But since only Honda makes Accords the real selectivity is $1/100 = 0.01$

Source: Guy Lohman

Statistics

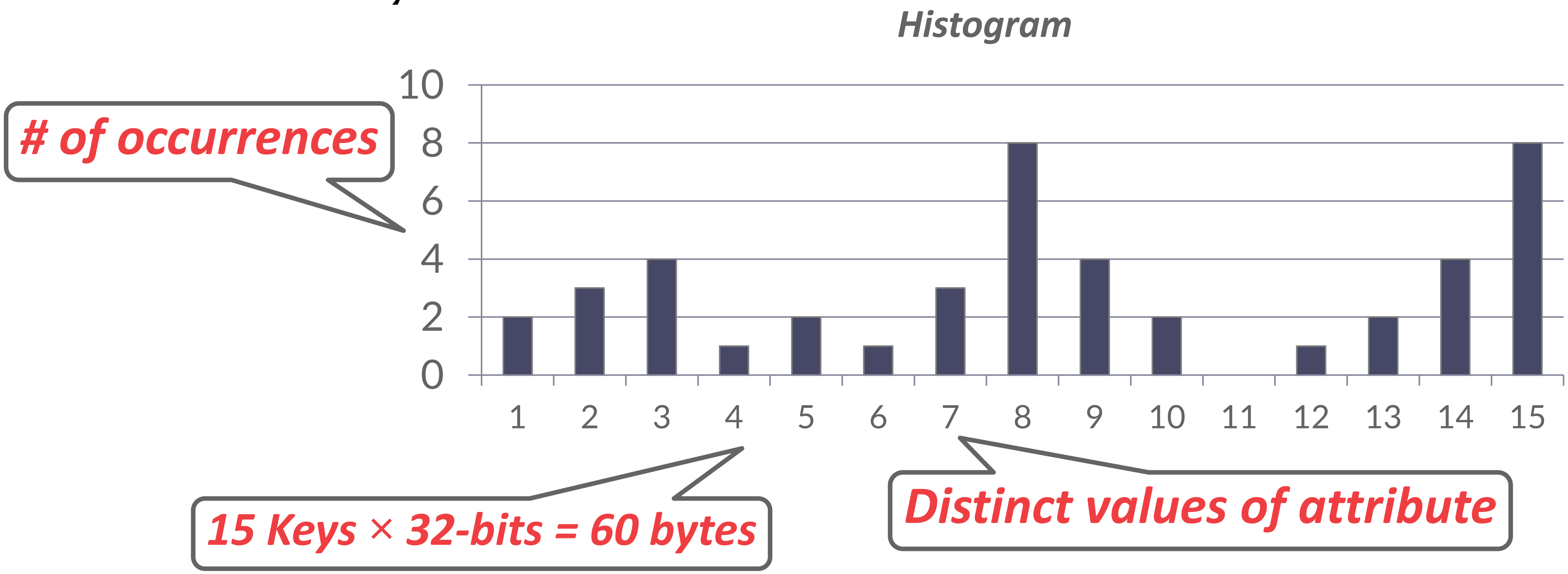
- **Choice #1: Histograms**
  - Maintain an occurrence count per value (or range of values) in a column.

- **Choice #2: Sketches**
  - Probabilistic data structure that gives an approximate count for a given value.

- **Choice #3: Sampling**
  - DBMS maintains a small subset of each table that it then uses to evaluate expressions to compute selectivity.

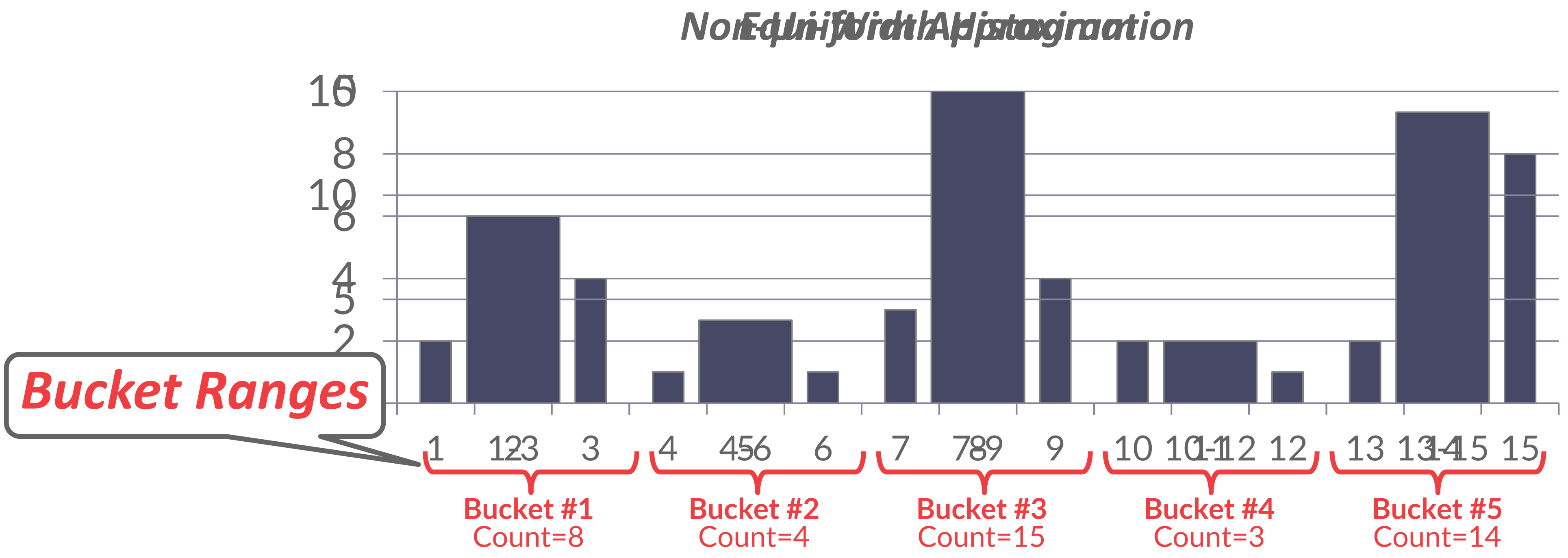
Histograms

- Our formulas are nice, but we assume that data values are uniformly distributed.



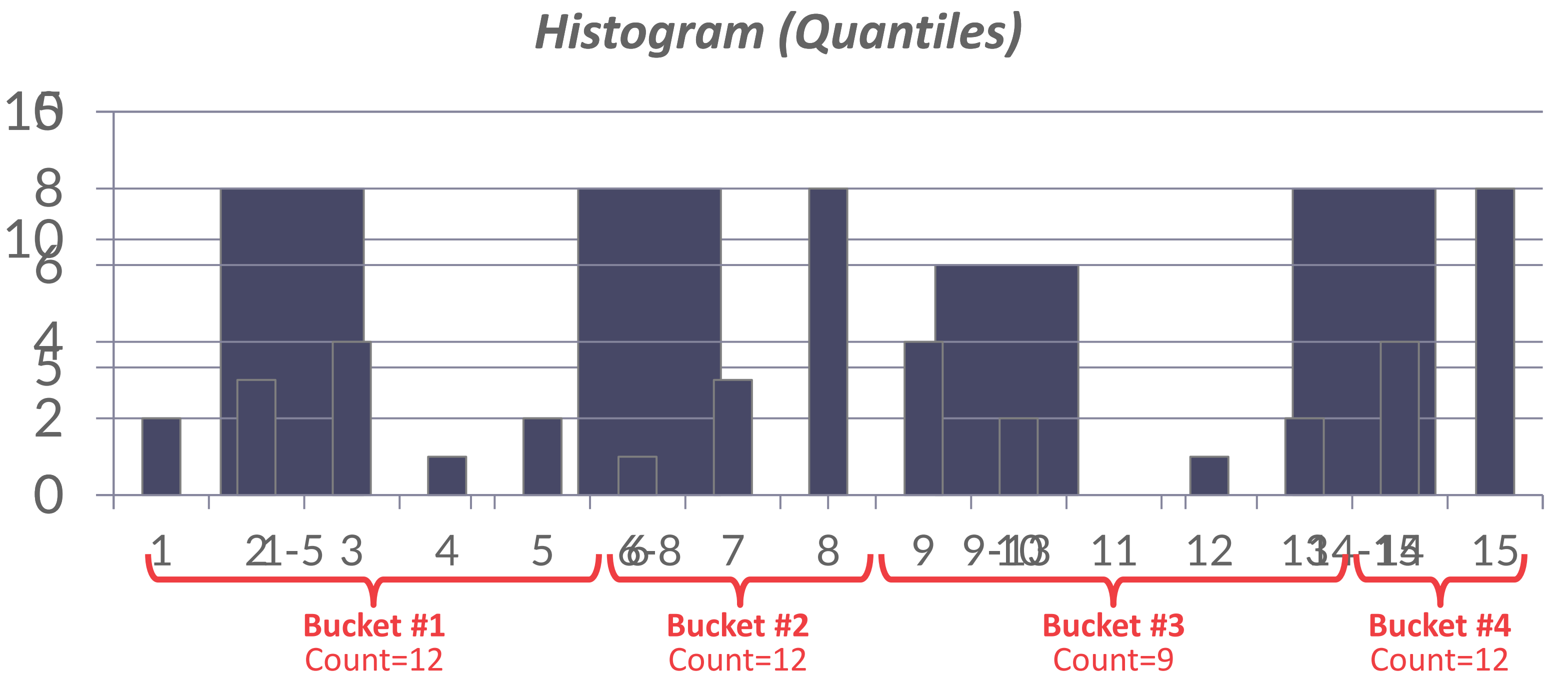
Equi-width Histogram

- Maintain counts for a group of values instead of each unique key. All buckets have the same width (i.e., same # of value).



Equi-depth Histogram

- Vary the width of buckets so that the total number of occurrences for each bucket is roughly the same.



Sketches

- Probabilistic data structures that generate approximate statistics about a data set.
- Cost-model can replace histograms with sketches to improve its selectivity estimate accuracy.

- Most common examples:
    - Count-Min Sketch (1988): Approximate frequency count of elements in a set.
    - HyperLogLog (2007): Approximate the number of distinct elements in a set.

Sampling

- Modern DBMSs also collect samples from tables to estimate selectivities.

- Update samples when the underlying tables changes significantly.

```
SELECT AVG(age)
  FROM people
 WHERE age > 50
```

id	name	age	status
1001	Obama	61	Rested
1002	Kanye	45	Weird
1003	Tupac	25	Dead
1004	Bieber	28	Crunk
1005	Andy	41	Illin
1006	TigerKin g	59	Jailed

1 billion tuples

Table Sample

1001	Obama	61	Rested
1003	Tupac	25	Dead
1005	Andy	41	Illin

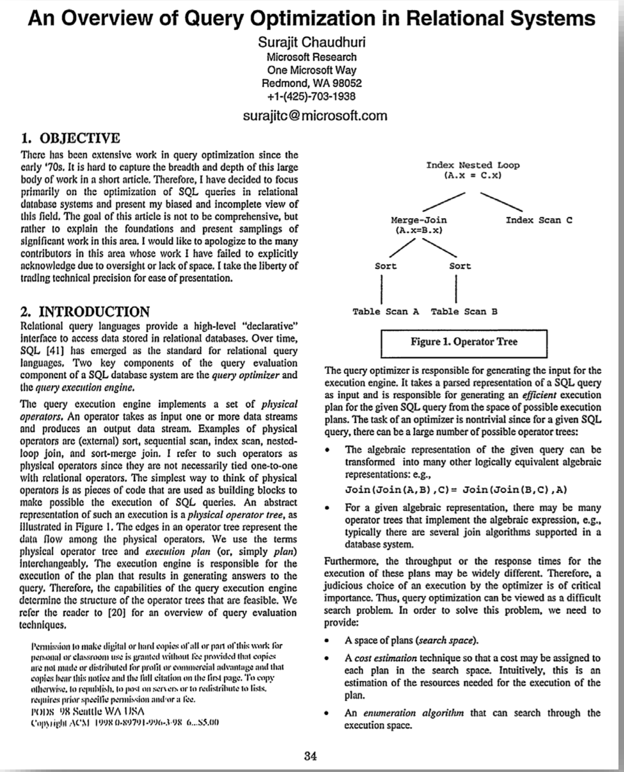
sel(age>50) = 1/3

Conclusion

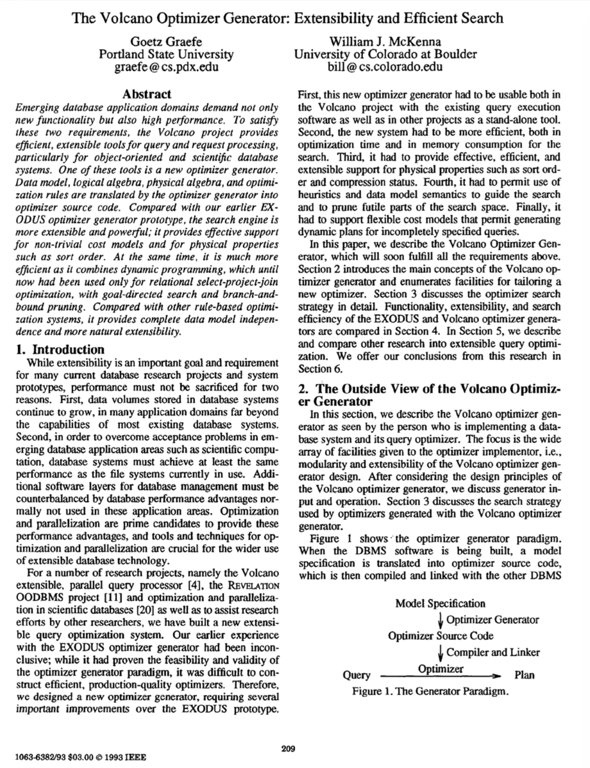
- Query optimization is critical for a database system.
- SQL -> logical plan -> physical plan.
- Flatten queries before going to the optimization part. Expression handling is also important.
- QO enumeration can be bottom-up or top-down.
- Need to cost each plan, so need cost-estimation methods.

Bonus

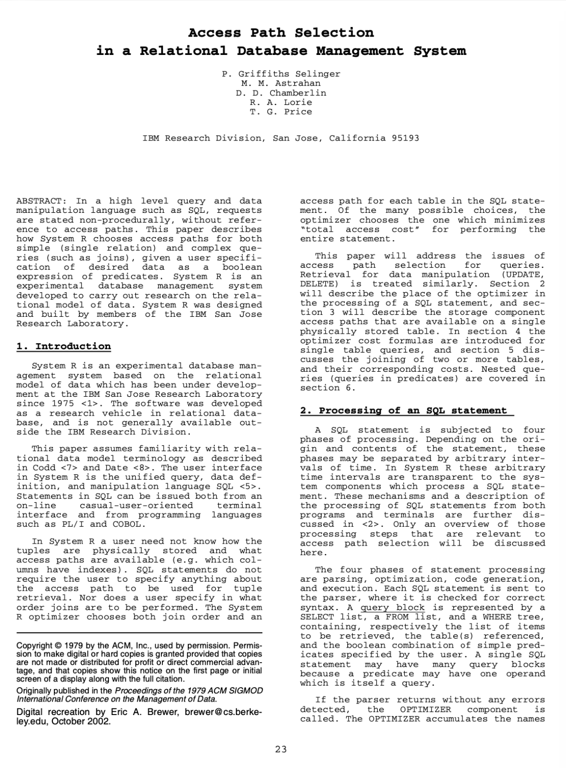
Essential Query Optimization papers



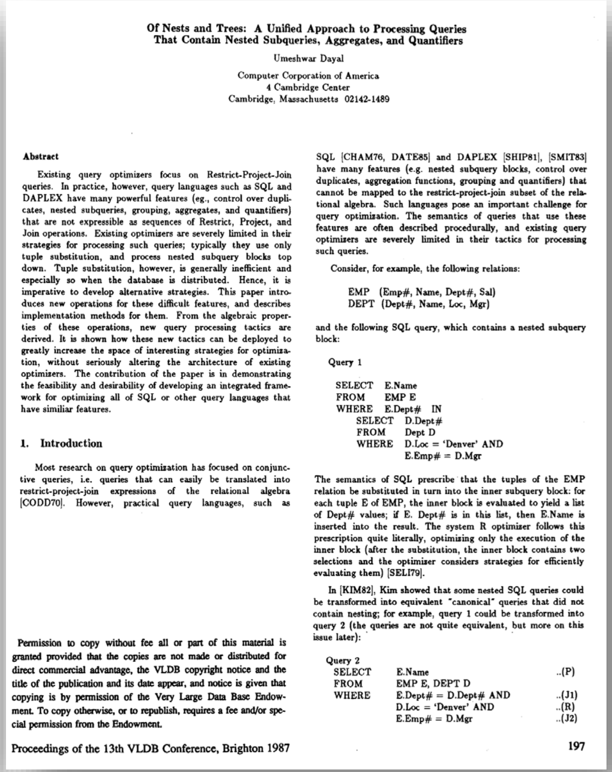
Surajit Chaudhuri: An Overview of Query Optimization in Relational Systems. PODS 1998: 34-43



Goetz Graefe, William J. McKenna: The Volcano Optimizer Generator: Extensibility and Efficient Search. ICDE 1993: 209-218



Patricia G. Selinger, Morton M. Astrahan, Donald D. Chamberlin, Raymond A. Lorie, Thomas G. Price: Access Path Selection in a Relational Database Management System. SIGMOD Conference 1979: 23-34



Umeshwar Dayal: Of Nests and Trees: A Unified Approach to Processing Queries That Contain Nested Subqueries, Aggregates, and Quantifiers. VLDB 1987: 197-208

Bonus

Suggestions if you are going to build a QO

- **Rule 1: Read lots of papers, especially from the 80s & 90s.**
  - Expect new combinations, only partially new core inventions.
- **Rule 2: Early on, test various workloads on the QO.**
  - QOs harden over time as they "see" new workloads. Let them see more ASAP.
- **Rule 3: Throw away the initial one (or two) and start anew.**
  - The hard part is going to be nitty-gritty details like data structures and pointers to shared objects; e.g., the list of predicates and the query graph structure, ... You will NOT get this right in the first pass. Don't try to patch; be prepared to rewrite.

Next Lecture

- DB Design
  - ER-diagrams, FDs