

Database System Internals Architecture

Paul G. Allen School of Computer Science and Engineering
University of Washington, Seattle

Announcements

- Lab 1 part 1 is due on Monday, April 4
 - Lab 1 in full is due on April 13
 - “git pull upstream master” before building
 - Remember to git commit and git push often!
 - Thursday’s sections: SimpleDB repo and structure
- HW1 is due on April 8
 - Submit via gradescope
- 544M first paper review is due April 15
 - Submit via gradescope
 - (Not hard deadline)

Important Note

- Lectures show principles
- Homeworks + Quizzes test the principles

- You need to think through what you will actually implement in SimpleDB!
 - Try to implement the simplest solutions

- If you are confused, let us know!

- SimpleDB not designed to be bullet-proof software

What we already know...

- **Database** = collection of related files
- **DBMS** = program that manages the database

What we already know...

- **Data models**: relational, semi-structured (XML), graph (RDF), key-value pairs
- **Relational model**: defines only the logical model, and does not define a physical storage of the data

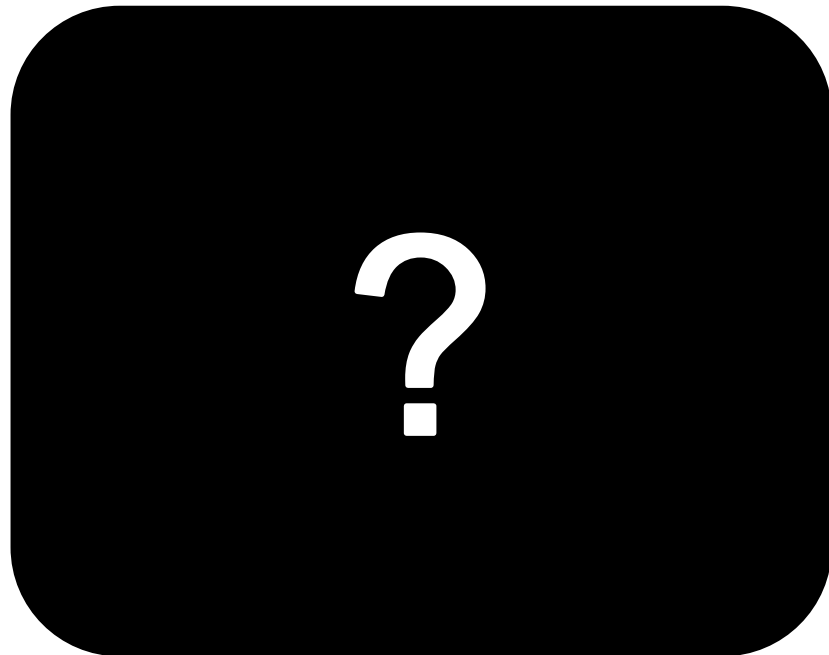
What we already know...

Relational Query Language:

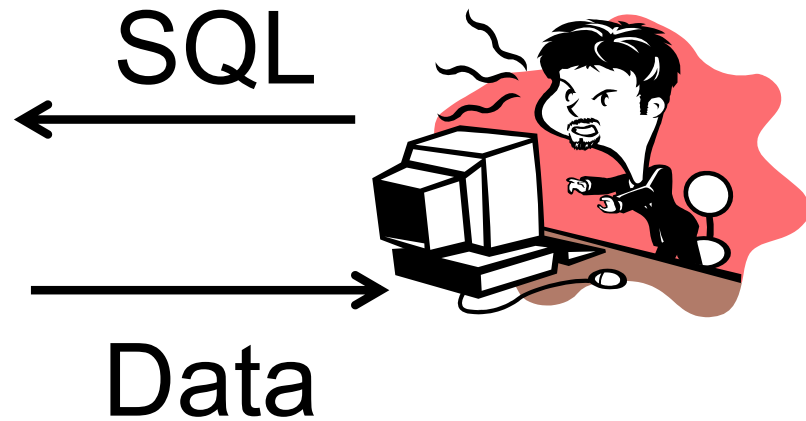
- **Set-at-a-time**: instead of tuple-at-a-time
- **Declarative**: user says what they want and not how to get it
- **Query optimizer**: from *what* to *how*

How to Implement an RDBMS?

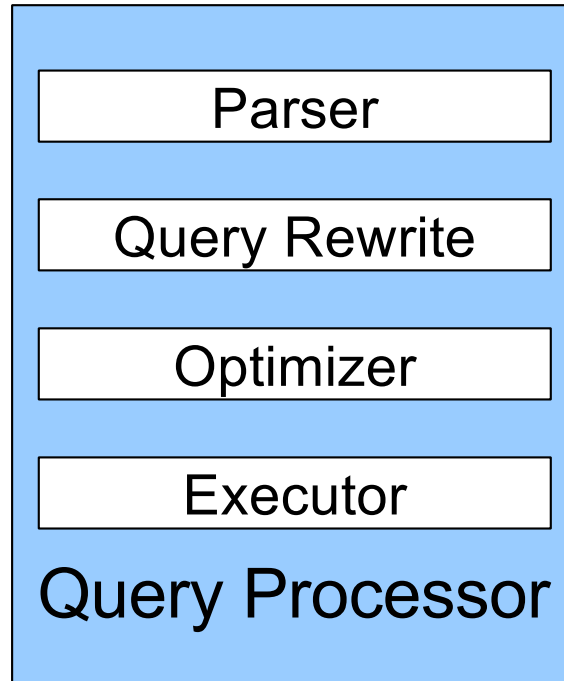
DBMS



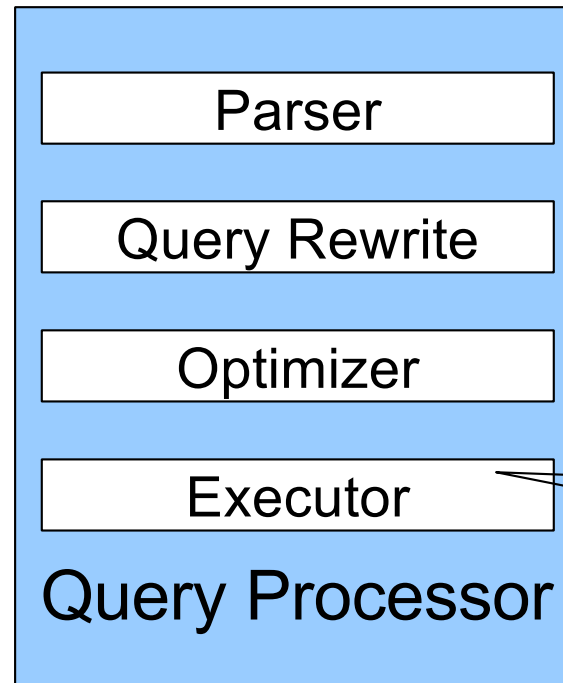
Key challenge: Achieve high performance on large databases!



DBMS Architecture

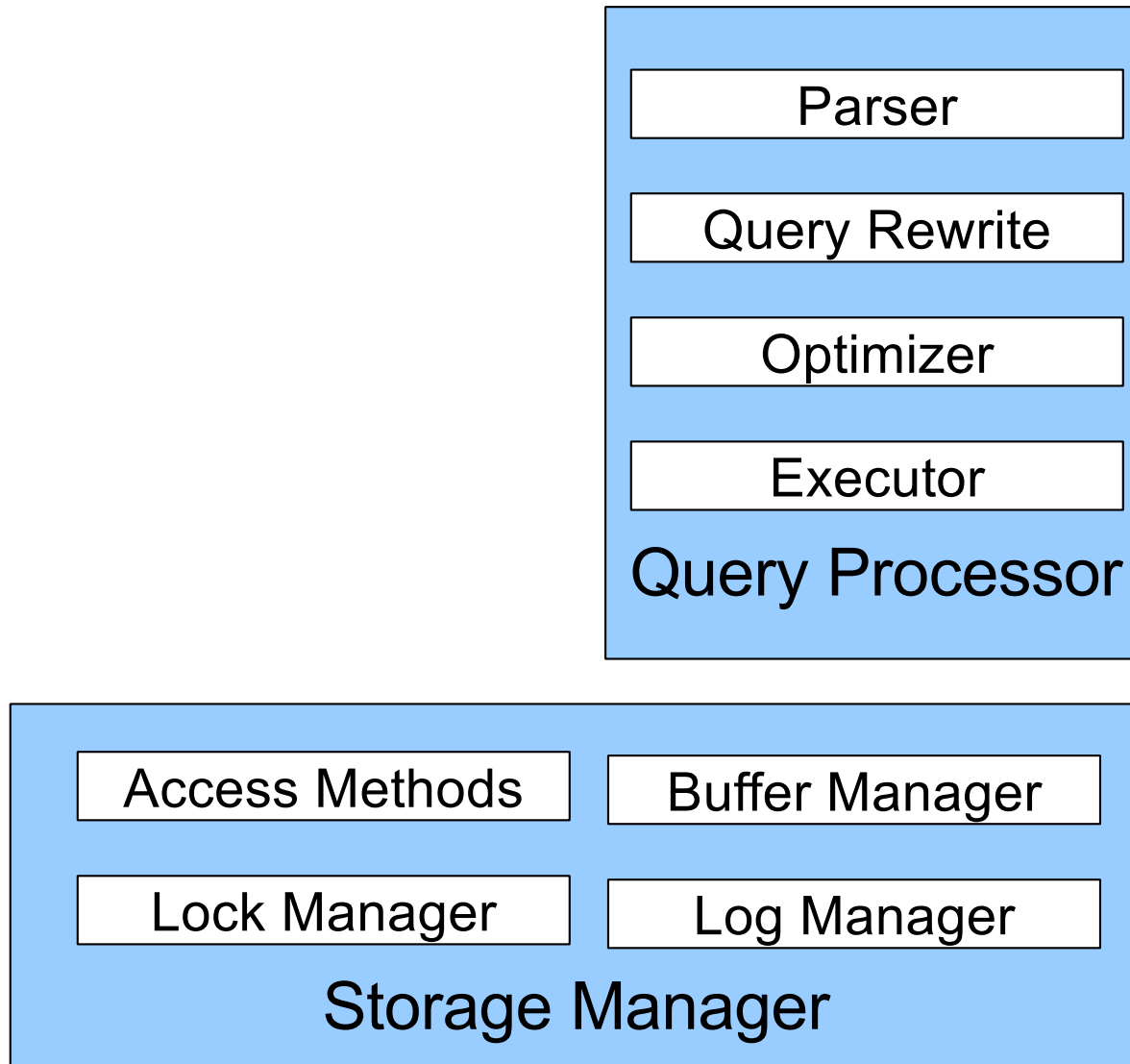


DBMS Architecture

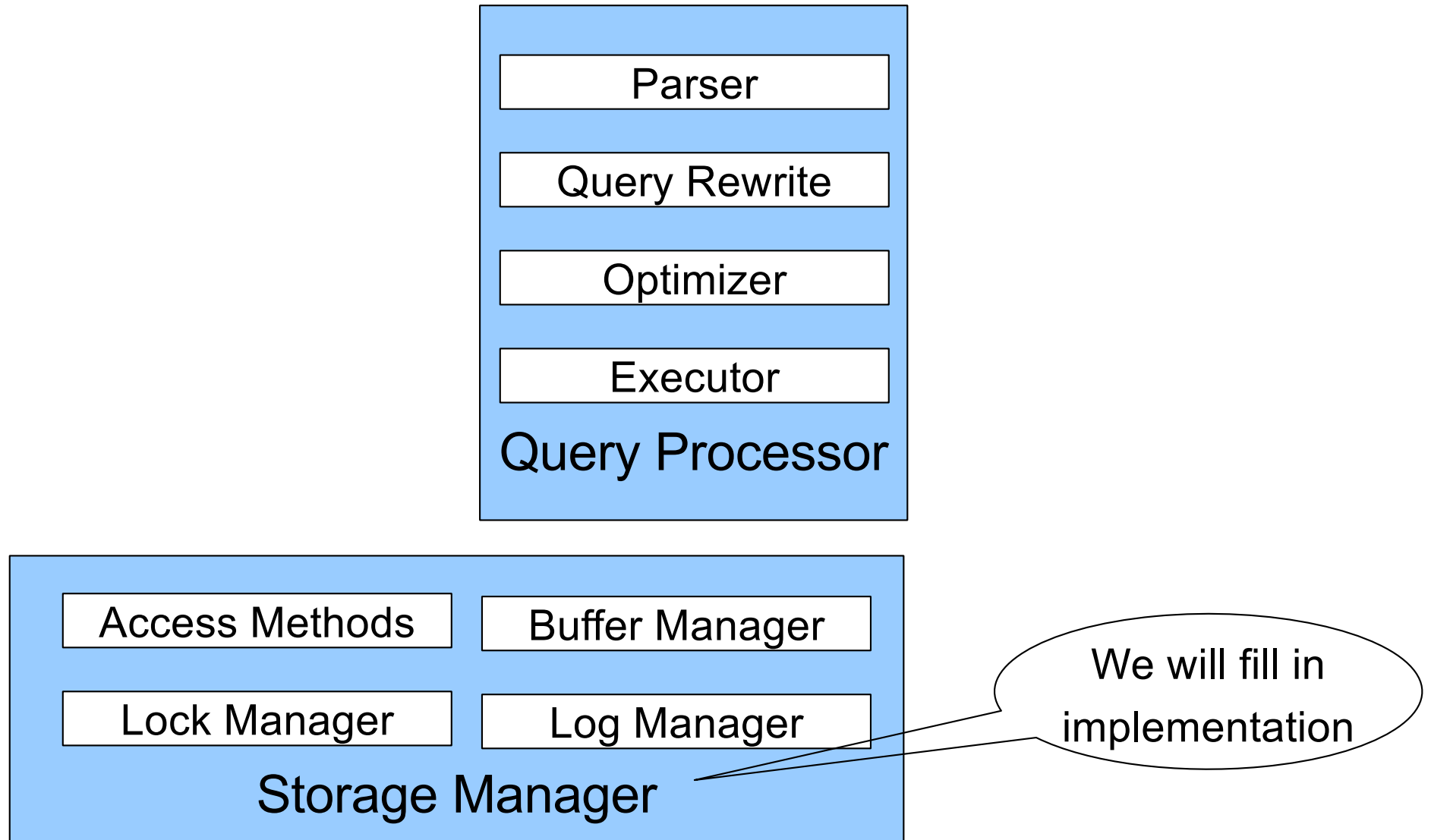


We will fill in implementation

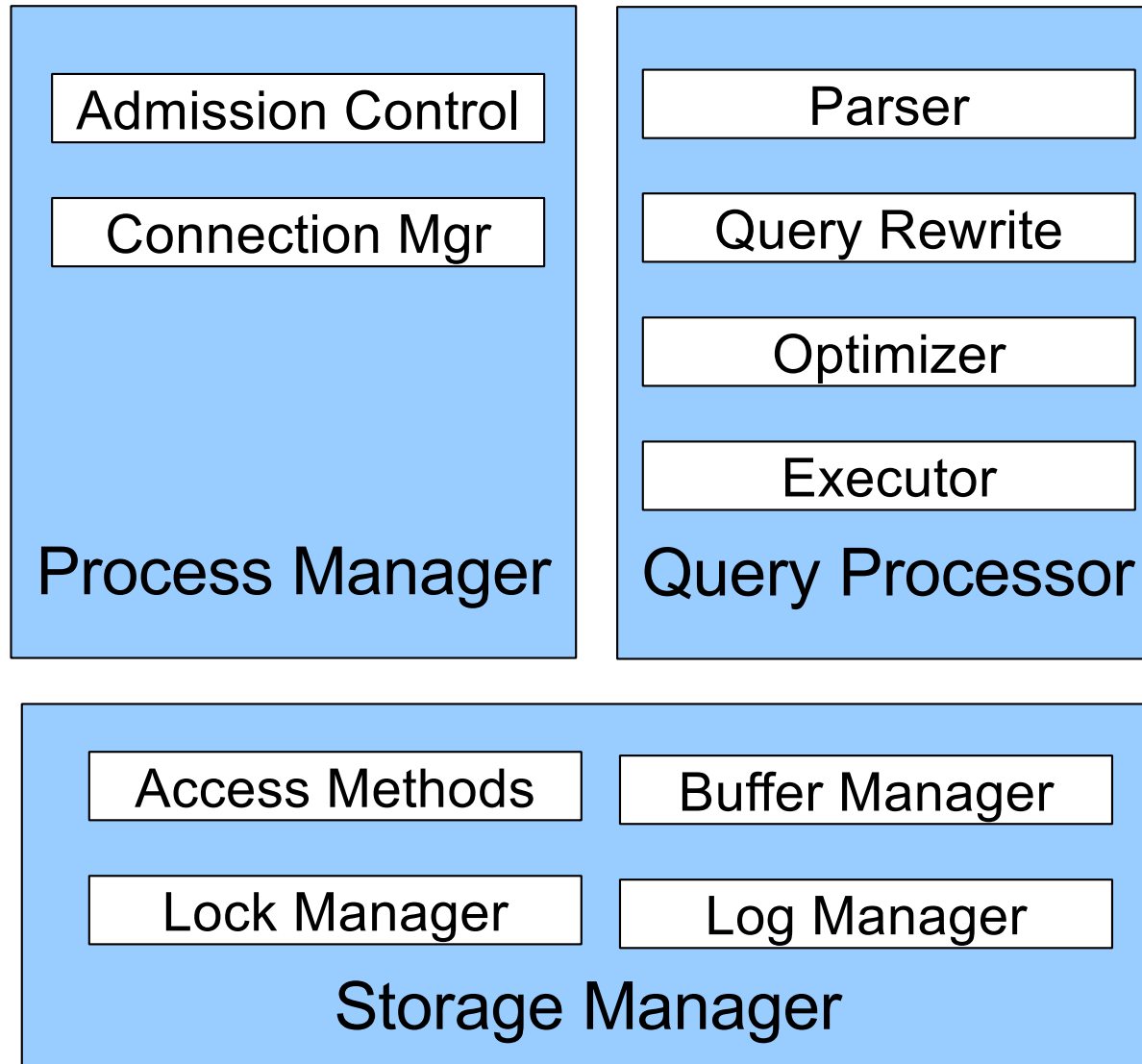
DBMS Architecture



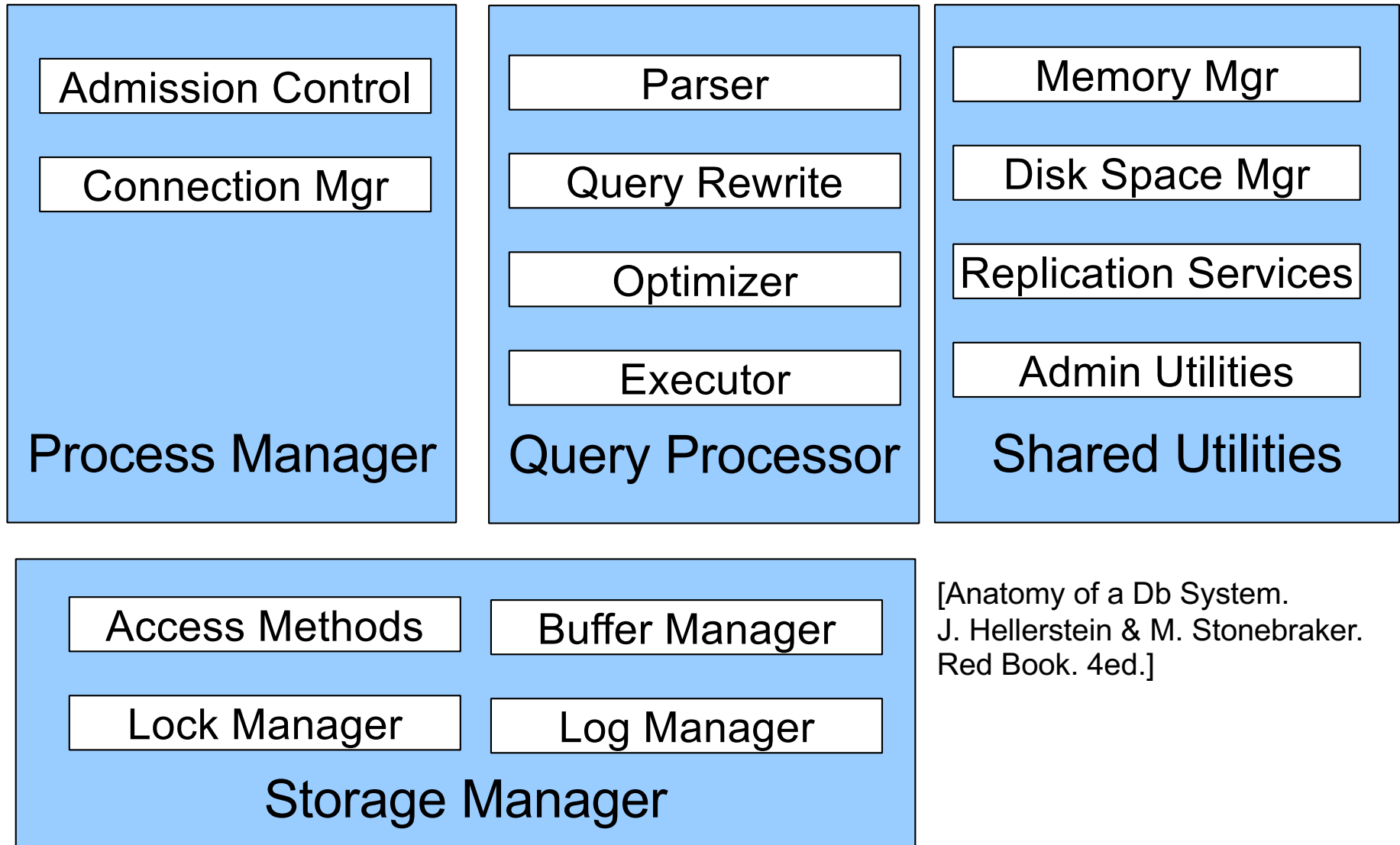
DBMS Architecture



DBMS Architecture



DBMS Architecture



[Anatomy of a Db System.
J. Hellerstein & M. Stonebraker.
Red Book. 4ed.]

Goal for Today

Overview of query execution

Overview of storage manager

Query Processor

Example Database Schema

Supplier (sno, sname, scity, sstate)

Part (pno, pname, psize, pcolor)

Supplies (sno, pno, price)

View: Suppliers in Seattle

```
CREATE VIEW NearbySupp AS
SELECT sno, sname
FROM Supplier
WHERE scity='Seattle' AND sstate='WA'
```


Example Query

```
Supplier (sno, sname, scity, sstate)  
Part (pno, pname, psize, pcolor)  
Supplies (sno, pno, price)
```

- Find the names of all suppliers in Seattle who supply part number 2

```
SELECT sname  
FROM NearbySupp  
WHERE sno IN ( SELECT sno  
                FROM Supplies  
                WHERE pno = 2 )
```

Query Processor

▪ Step 1: Parser

- Parses query into an internal format
- Performs various checks using **catalog**

▪ Step 2: Query rewrite

- View rewriting, flattening, etc.

Rewritten Version of Our Query

```
Supplier (sno, sname, scity, sstate)  
Part (pno, pname, psize, pcolor)  
Supplies (sno, pno, price)
```

Original query:

```
SELECT sname  
FROM NearbySupp  
WHERE sno IN ( SELECT sno  
                FROM Supplies  
                WHERE pno = 2 )
```

Rewritten query (expanding NearbySupp view):

```
SELECT S.sname  
FROM Supplier S, Supplies U  
WHERE S.scity='Seattle' AND S.sstate='WA'  
AND S.sno = U.sno  
AND U.pno = 2;
```

Query Processor

▪ Step 3: Optimizer

- Find an efficient query plan for executing the query
- **A query plan is**
 - **Logical:** An extended relational algebra tree
 - **Physical:** With additional annotations at each node
 - Access method to use for each relation
 - Implementation to use for each relational operator

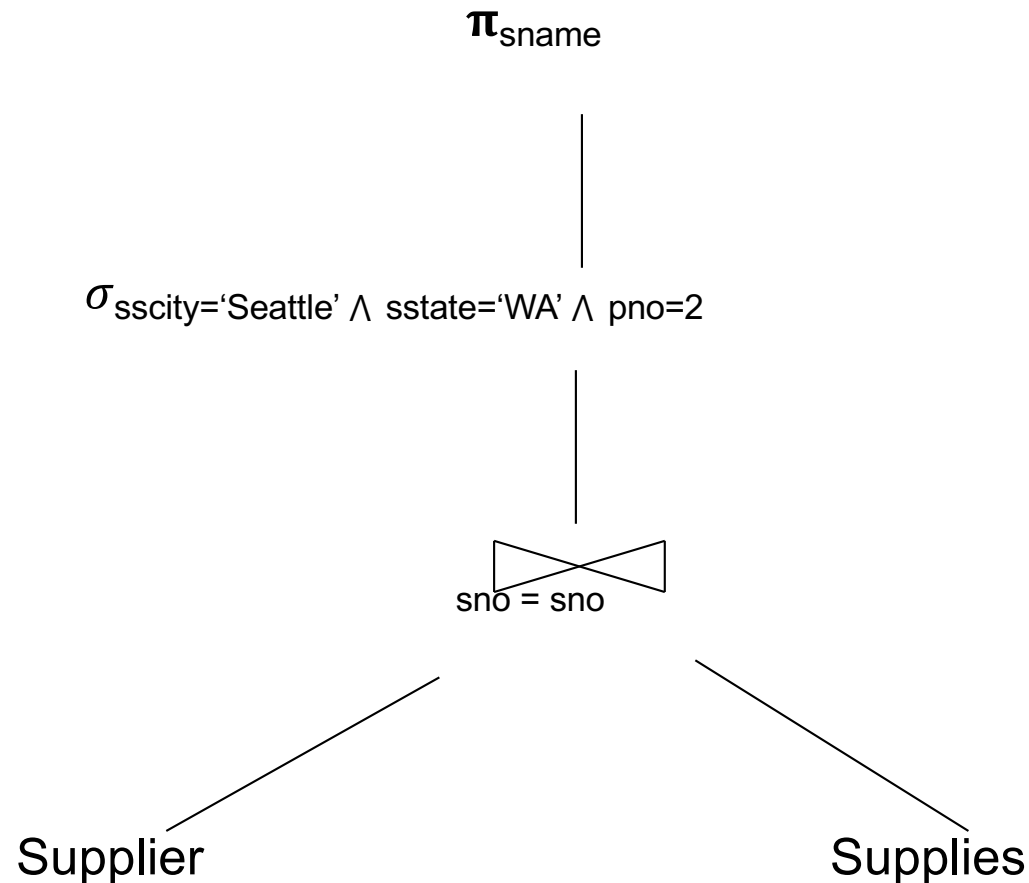
▪ Step 4: Executor

- Actually executes the physical plan

Logical Query Plan

```
SELECT S.sname
FROM Supplier S, Supplies U
WHERE
S.scity='Seattle'
AND S.sstate='WA'
AND S.sno = U.sno
AND U.pno = 2;
```

Supplier (sno, sname, scity, sstate)
Part (pno, pname, psize, pcolor)
Supplies (sno, pno, price)



Physical Query Plan

- Logical query plan with extra annotations
- **Implementation choice** for each operator
- **Access path selection** for each relation
 - Bottom of tree = read from disk
 - Use a file scan or use an index

Physical Query Plan

“on the fly” means the tuples are processed one-by-one as they pass through the operator

Supplier (sno, sname, scity, sstate)
Part (pno, pname, psize, pcolor)
Supplies (sno, pno, price)

(On the fly)

π_{sname}

(On the fly)

$\sigma_{\text{scity}='Seattle' \wedge \text{sstate}='WA' \wedge \text{pno}=2}$

(Nested loop)

sno = sno

Suppliers

(File scan)

Supplies

(File scan)

Query Executor

Tuples in SimpleDB

`Tuple.java` describes a row object in SimpleDB

- Rows are the objects passed through the database
- In the same way we conceptualize RA and a series of transformations to rows, so does it work in database

Iterator Interface

- Each **operator implements Operator.java**

open()

- Initializes operator state
- Sets parameters such as selection predicate

next()

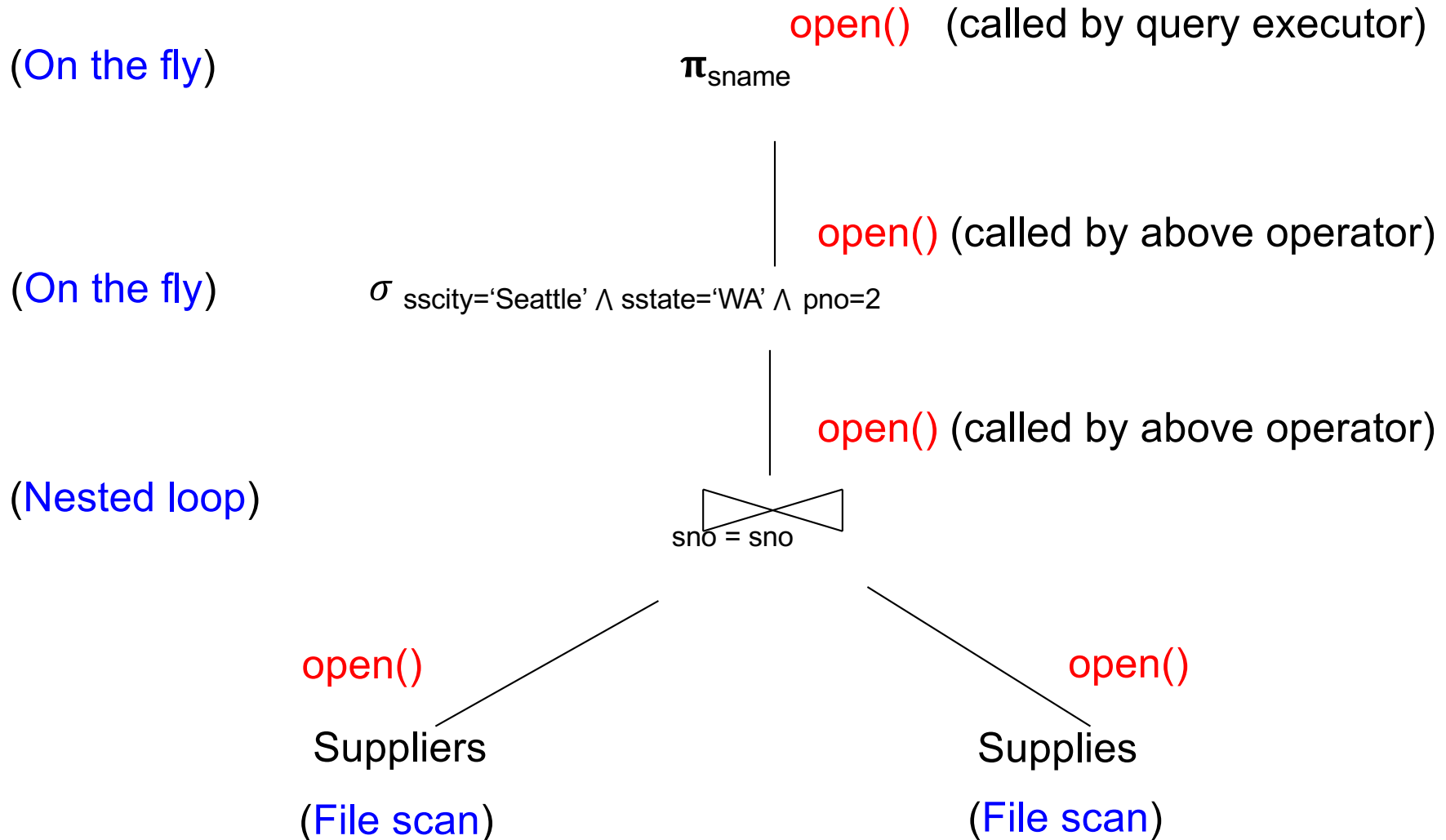
- **Returns a Tuple!**
- Operator invokes next() recursively on its inputs
- Performs processing and produces an output tuple

close() clean-up state

- Operators also have reference to their **child** operator in the query plan
can call `child.open()` , `child.next()` etc..

Query Execution

Supplier (sno, sname, scity, sstate)
Part (pno, pname, psize, pcolor)
Supplies (sno, pno, price)



Query Execution

Supplier (sno, sname, scity, sstate)
Part (pno, pname, psize, pcolor)
Supplies (sno, pno, price)

(On the fly)

π_{sname} next()

pull-based execution

(On the fly)

$\sigma_{\text{scity}='Seattle' \wedge \text{sstate}='WA' \wedge \text{pno}=2}$ next()

(Nested loop)

next()
sno = sno

next()

Suppliers

(File scan)

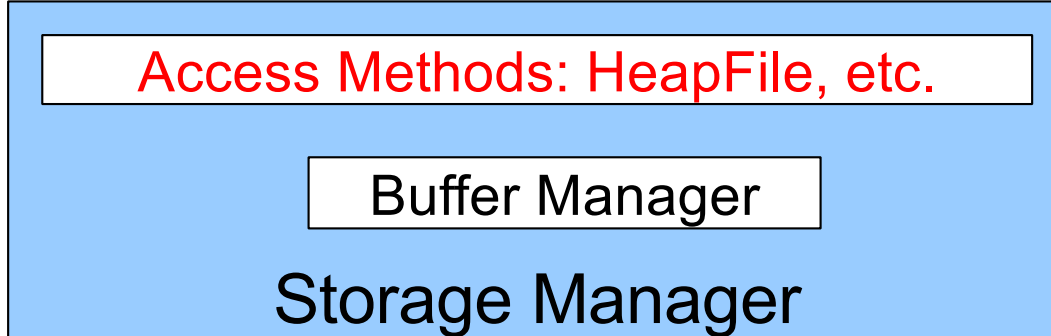
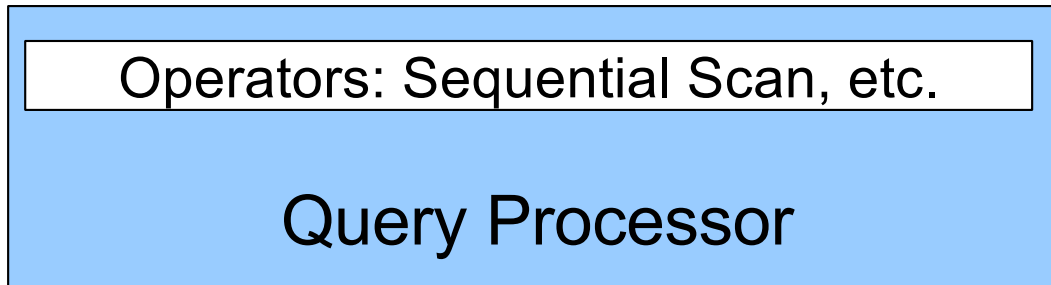
next()

Supplies

(File scan)

Storage Manager

Access Methods



- **Operators:** Process data
- **Access methods:** Organize data to support fast access to desired subsets of records
- **Buffer manager:** Caches data in memory. Reads/writes data to/from disk as needed
- **Disk-space manager:** Allocates space on disk for files/access methods

Disk Storage

- Can only read **1 block per read operation**

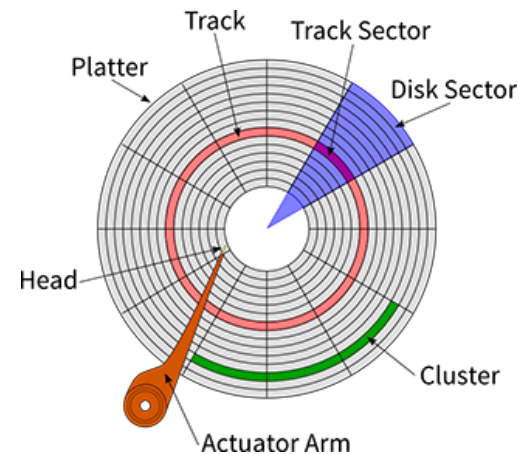
- Usually 512B to 4kB

- One blocks contains some Tuples

444	Ryan	G20
344	Ryan	134
544	Dan	134

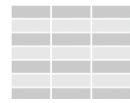
- **Sequential disk reads are faster than random ones**

- Cost ~1-2% random scan = full sequential scan



Buffer Manager (BufferPool in SimpleDB)

Page requests from higher-level code



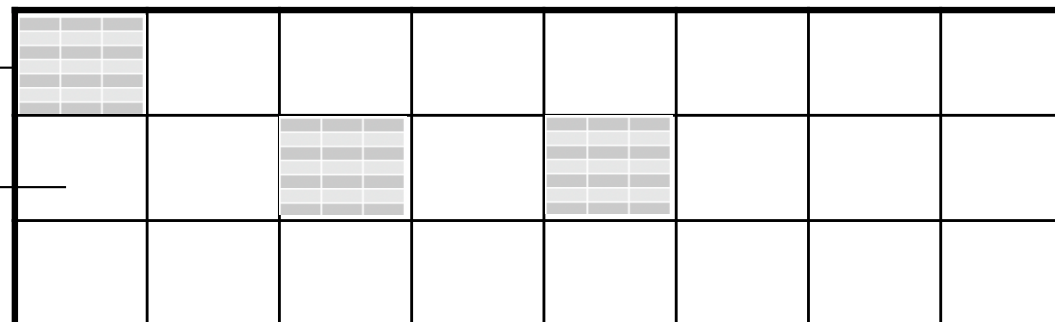
Access methods

Buffer pool manager

Buffer pool

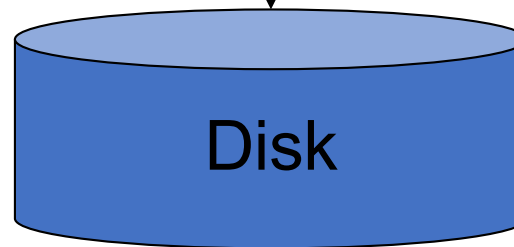
Disk page

Free frame



Main memory

Disk is a collection of blocks



1 page corresponds to 1 disk block

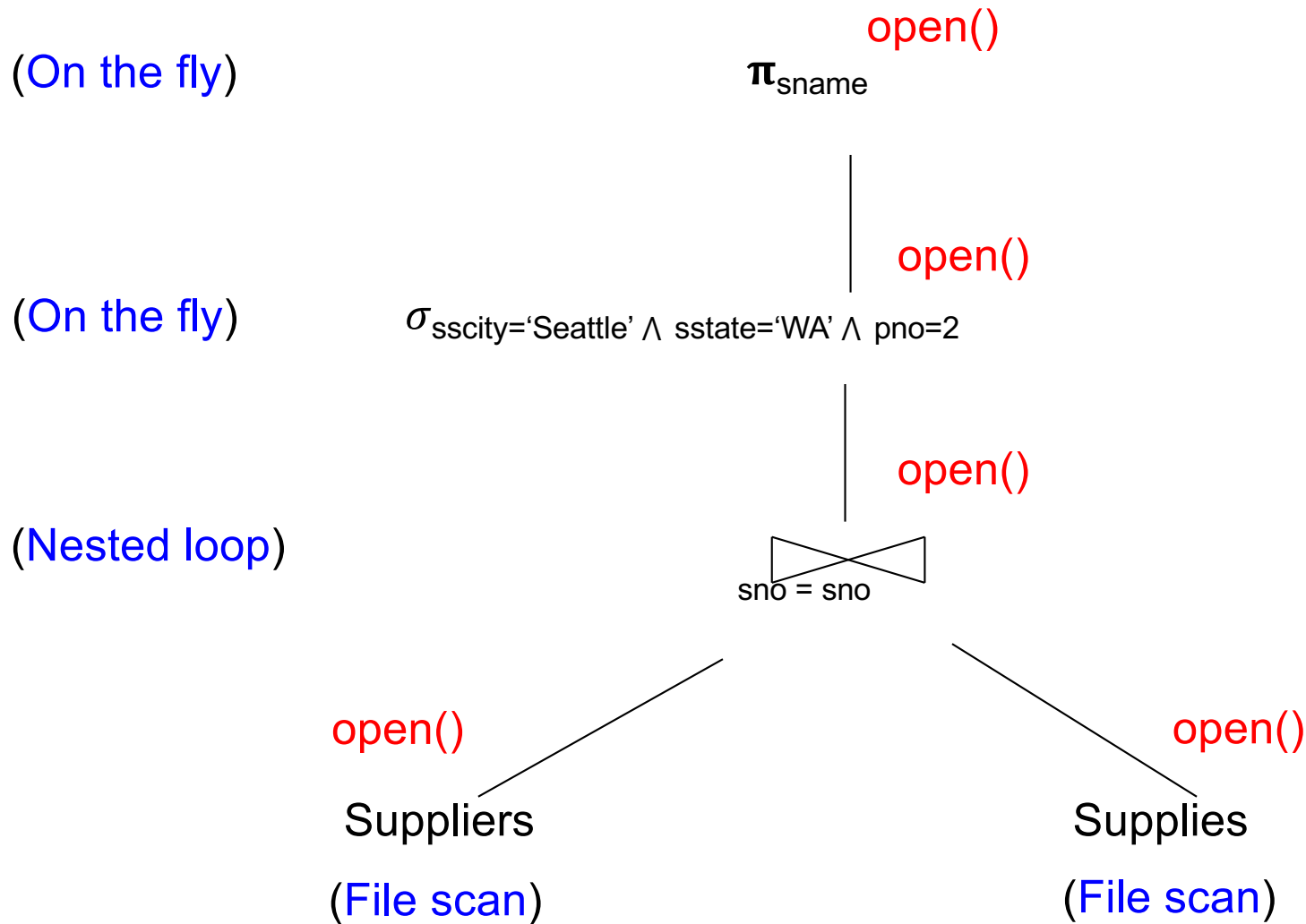
Buffer Manager

- Brings pages in from memory and caches them
- Eviction policies
 - Random page (ok for SimpleDB)
 - Least-recently used
 - The “clock” algorithm (see book)
- Keeps track of which **pages are dirty**
 - A dirty page has changes not reflected on disk
 - Implementation: Each page includes a dirty bit

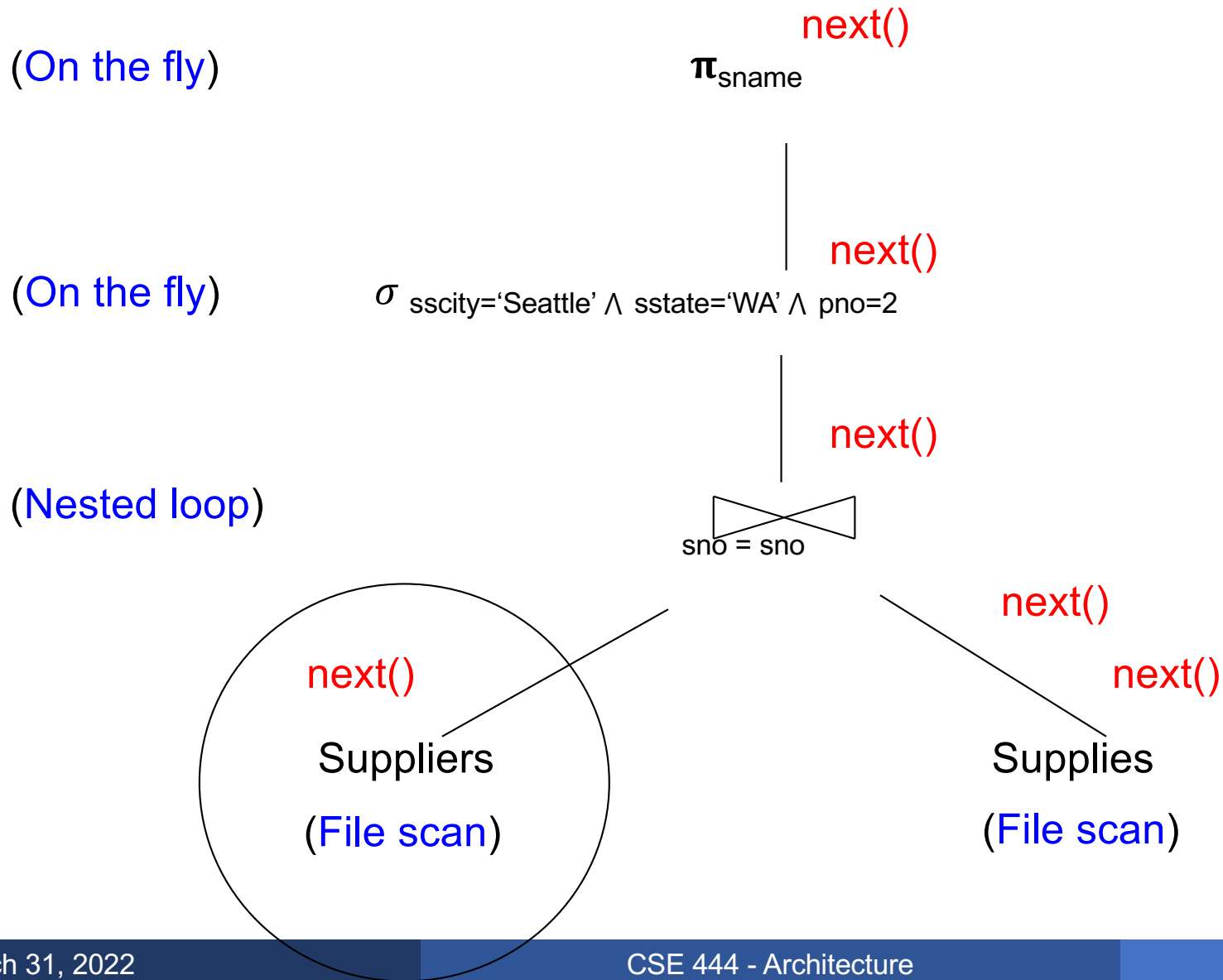
Access Methods

- A DBMS stores data on disk by breaking it into *pages*
 - A page is the size of a disk block.
 - A page is the unit of disk IO
- Buffer manager caches these pages in memory
- Access methods do the following:
 - They organize pages into collections called DB *files*
 - They organize data inside pages
 - They provide an API for operators to access data in these files
- Discussion:
 - OS vs DBMS files
 - OS vs DBMS buffer manager

Query Execution



Query Execution



Query Execution In SimpleDB

open()

next()

SeqScan

Operator at
bottom of plan

open()

next()

Heap File Access Method

In SimpleDB, SeqScan can
find HeapFile in Catalog

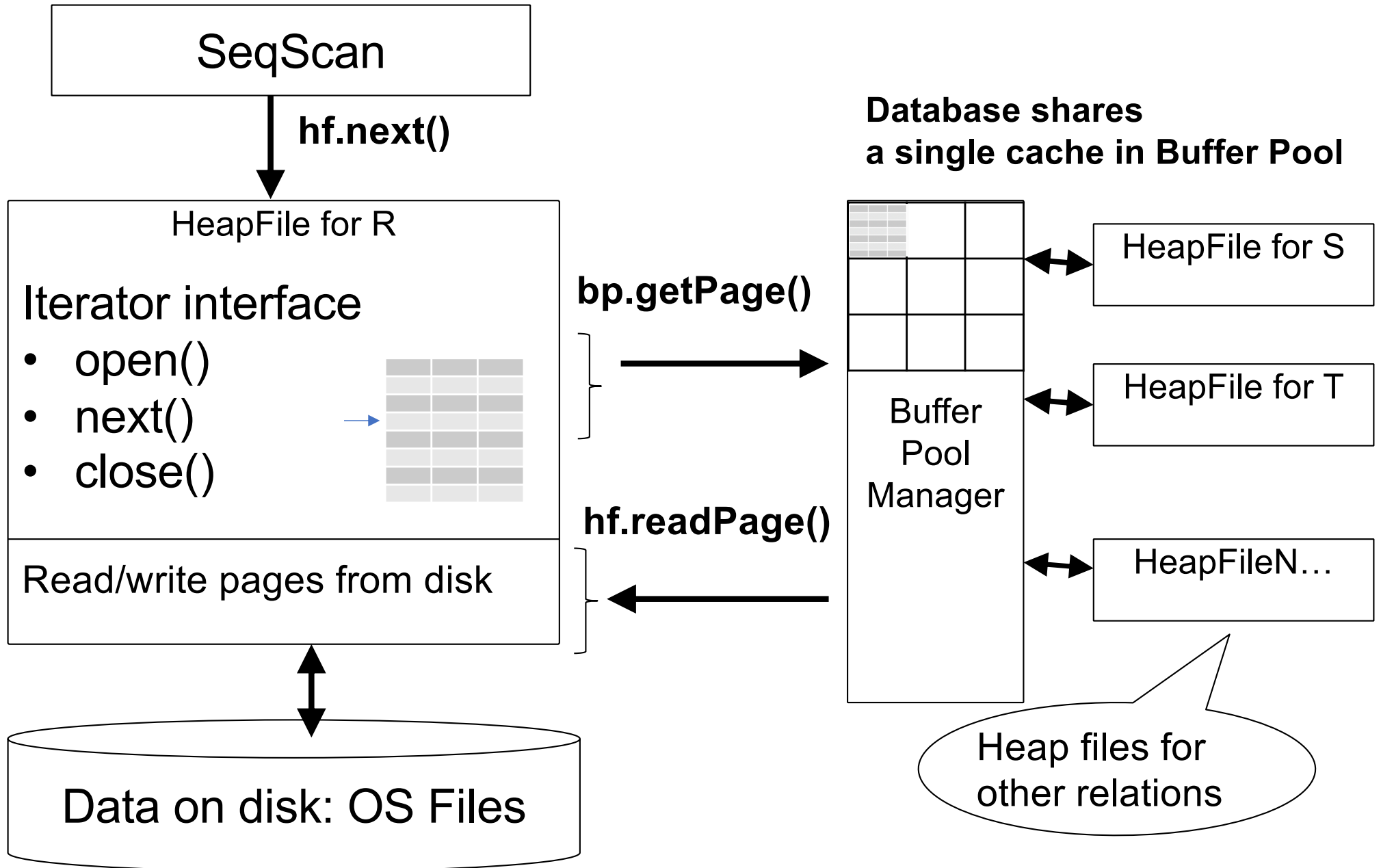
Offers iterator interface

- open()
- next()
- close()

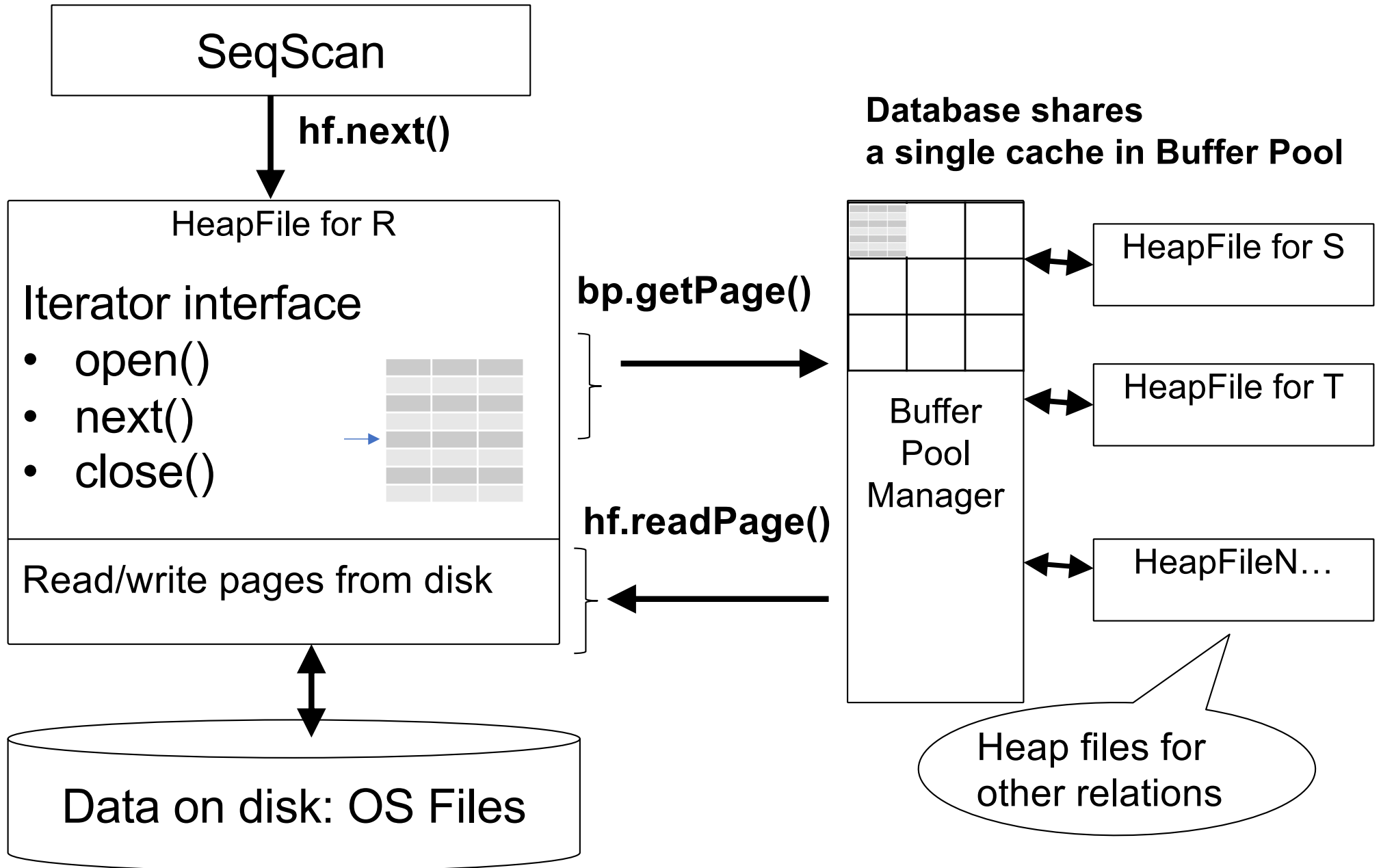
Knows how to read/write pages from disk

But if Heap File reads data
directly from disk, it will not
stay cached in Buffer Pool!

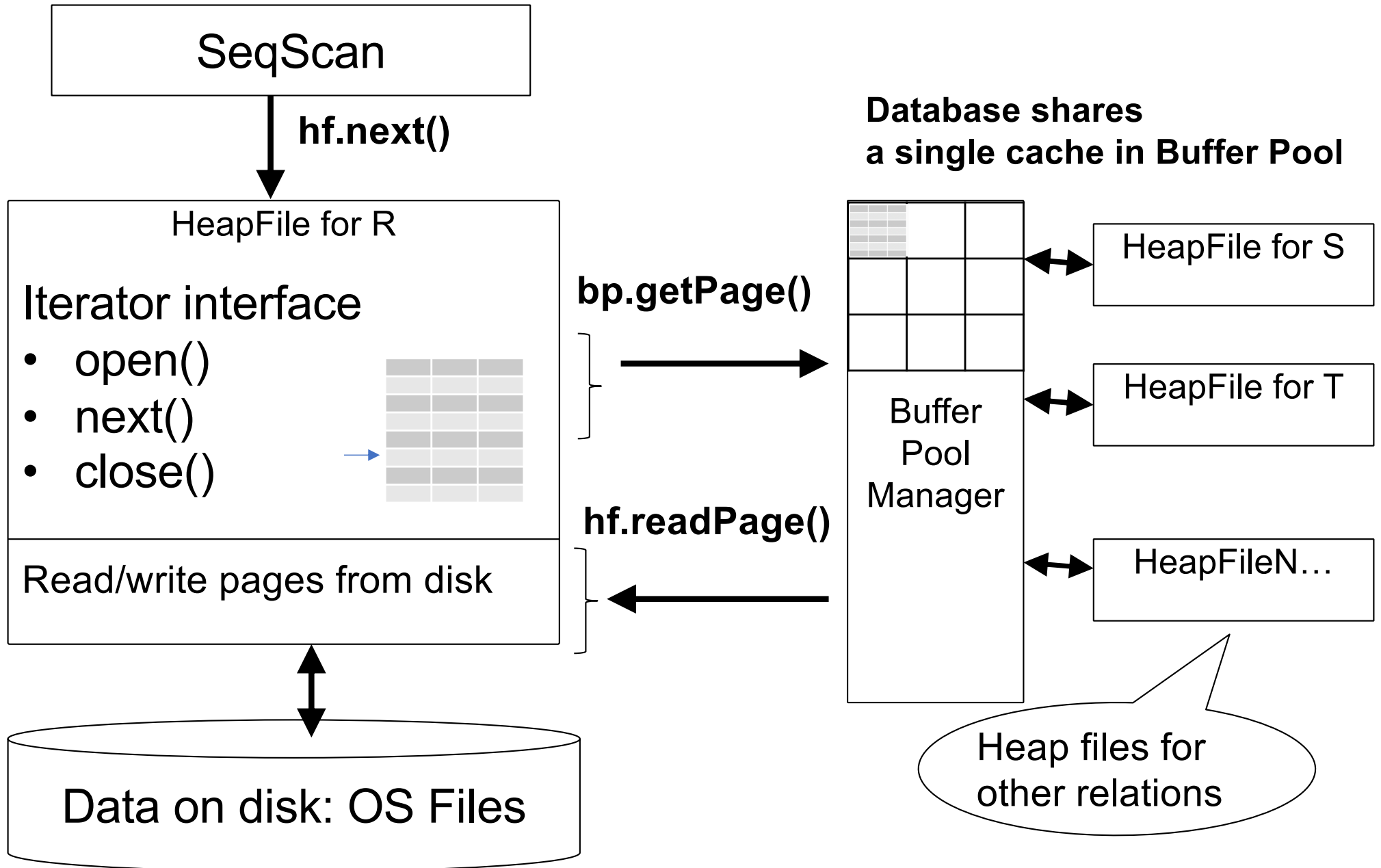
Query Execution In SimpleDB



Query Execution In SimpleDB



Query Execution In SimpleDB



HeapFile In SimpleDB

- Data is stored on disk in an OS file. HeapFile class knows how to “decode” its content
- Control flow:

SeqScan calls methods such as "iterate" on the HeapFile Access Method

During the iteration, the HeapFile object needs to call the BufferManager.getPage() method to ensure that necessary pages get loaded into memory.

The BufferManager will then call HeapFile .readPage()/writePage() page to actually read/write the page.