Database Storage — Part I
Homework #1 is due September 12th @ 11:59pm

Project #0 is due September 12th @ 11:59pm

Project #1 will be released on September 13th
OVERVIEW

We now understand what a database looks like at a logical level and how to write queries to read/write data (e.g., using SQL).

We will next learn how to build software that manages a database (i.e., a DBMS).
COURSE OUTLINE

Relational Databases
Storage
Execution
Concurrency Control
Recovery
Distributed Databases
Potpourri

Query Planning
Operator Execution
Access Methods
Buffer Pool Manager
Disk Manager
COURSE OUTLINE

Relational Databases
Storage
Execution
Concurrency Control
Recovery
Distributed Databases
Potpourri

Query Planning
Operator Execution
Access Methods
Buffer Pool Manager
Disk Manager
DISK-BASED ARCHITECTURE

The DBMS assumes that the primary storage location of the database is on non-volatile disk.

The DBMS's components manage the movement of data between non-volatile and volatile storage.
STORAGE HIERARCHY

- CPU Registers
- CPU Caches
- DRAM
- SSD
- HDD
- Network Storage
STORAGE HIERARCHY

- CPU Registers
- CPU Caches
- DRAM
- SSD
- HDD
- Network Storage

Faster, Smaller, Expensive
Slower, Larger, Cheaper
STORAGE HIERARCHY

- **CPU Registers**
  - Faster
  - Smaller
  - Expensive

- **CPU Caches**
  - Faster
  - Smaller
  - Expensive

- **DRAM**
  - Faster
  - Larger
  - Cheaper

- **SSD**
  - Faster
  - Smaller
  - Cheaper

- **HDD**
  - Slower
  - Larger
  - Cheaper

- **Network Storage**
  - Slower
  - Larger
  - Cheaper

**Volatile**
- Random Access
- Byte-Addressable

**Non-Volatile**
- Sequential Access
- Block-Addressable
STORAGE HIERARCHY

- CPU Registers
- CPU Caches
- DRAM
- SSD
- HDD
- Network Storage

Faster: CPU Registers
Smaller: CPU Registers
Expensive: CPU Registers
Slower: Network Storage
Larger: Network Storage
Cheaper: Network Storage
STORAGE HIERARCHY

- CPU Registers
- CPU Caches
- DRAM
- SSD
- HDD
- Network Storage

Faster, Smaller, Expensive
Slower, Larger, Cheaper
STORAGE HIERARCHY

- CPU Registers
- CPU Caches
- DRAM
- SSD
- HDD
- Network Storage

Memory levels:
- Faster
- Smaller
- Expensive

Disk levels:
- Slower
- Larger
- Cheaper
STORAGE HIERARCHY

- CPU Registers
- CPU Caches
- DRAM
- SSD
- HDD
- Network Storage

Faster
Smaller
Expensive

Slower
Larger
Cheaper
STORAGE HIERARCHY

- **CPU**
  - CPU Registers
  - CPU Caches

- **Memory**
  - DRAM

- **Disk**
  - SSD
    - Fast Network Storage
  - HDD
  - Network Storage

**Properties**

- Faster
- Smaller
- Expensive
- Slower
- Larger
- Cheaper

**Fast Network Storage**
STORAGE HIERARCHY

- CPU Registers
- CPU Caches
- DRAM
- Non-volatile Memory
- SSD
- Fast Network Storage
- HDD
- Network Storage

Faster
Smaller
Expensive

Slower
Larger
Cheaper
STORAGE HIERARCHY

- CPU Registers
- CPU Caches
- DRAM
- Non-volatile Memory
- SSD
- Fast Network Storage
- HDD
- Network Storage

**Memory**
- Faster
- Smaller
- Expensive

**Disk**
- Larger
- Cheaper

Non-Volatile Memory Database Management Systems
Joy Arulraj
Andrew Pavlo
STORAGE HIERARCHY

Memory

Disk

CPU

CPU Registers

CPU Caches

DRAM

SSD

HDD

Network Storage

Faster

Smaller

Expensive

Slower

Larger

Cheaper

Non-Volatile Memory Database Management Systems

Joy Arulraj
Andrew Pavlo

CMU-DB 15-445/645 (Fall 2021)
## ACCESS TIMES

<table>
<thead>
<tr>
<th>Storage Type</th>
<th>Access Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Cache Ref</td>
<td>0.5 ns</td>
</tr>
<tr>
<td>L2 Cache Ref</td>
<td>7 ns</td>
</tr>
<tr>
<td>DRAM</td>
<td>100 ns</td>
</tr>
<tr>
<td>SSD</td>
<td>150,000 ns</td>
</tr>
<tr>
<td>HDD</td>
<td>10,000,000 ns</td>
</tr>
<tr>
<td>Network Storage</td>
<td>~30,000,000 ns</td>
</tr>
<tr>
<td>Tape Archives</td>
<td>1,000,000,000 ns</td>
</tr>
</tbody>
</table>

(Source)
## ACCESS TIMES

<table>
<thead>
<tr>
<th>Storage Type</th>
<th>Access Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Cache Ref</td>
<td>0.5 ns</td>
</tr>
<tr>
<td>L2 Cache Ref</td>
<td>7 ns</td>
</tr>
<tr>
<td>DRAM</td>
<td>100 ns</td>
</tr>
<tr>
<td>SSD</td>
<td>150,000 ns</td>
</tr>
<tr>
<td>HDD</td>
<td>10,000,000 ns</td>
</tr>
<tr>
<td>Network Storage</td>
<td>~30,000,000 ns</td>
</tr>
<tr>
<td>Tape Archives</td>
<td>1,000,000,000 ns</td>
</tr>
<tr>
<td></td>
<td><strong>0.5 sec</strong></td>
</tr>
<tr>
<td></td>
<td><strong>7 sec</strong></td>
</tr>
<tr>
<td></td>
<td><strong>100 sec</strong></td>
</tr>
<tr>
<td></td>
<td><strong>1.7 days</strong></td>
</tr>
<tr>
<td></td>
<td><strong>16.5 weeks</strong></td>
</tr>
<tr>
<td></td>
<td><strong>11.4 months</strong></td>
</tr>
<tr>
<td></td>
<td><strong>31.7 years</strong></td>
</tr>
</tbody>
</table>

[Source]
SEQUENTIAL VS. RANDOM ACCESS

Random access on non-volatile storage is usually much slower than sequential access.

DBMS will want to maximize sequential access.
→ Algorithms try to reduce number of writes to random pages so that data is stored in contiguous blocks.
→ Allocating multiple pages at the same time is called an extent.
SYSTEM DESIGN GOALS

Allow the DBMS to manage databases that exceed the amount of memory available.

Reading/writing to disk is expensive, so it must be managed carefully to avoid large stalls and performance degradation.

Random access on disk is usually much slower than sequential access, so the DBMS will want to maximize sequential access.
DISK-ORIENTED DBMS
DISK-ORIENTED DBMS
DISK-ORIENTED DBMS
DISK-ORIENTED DBMS

Memory

Buffer Pool

Execution Engine

Disk

Database File

Directory

Page Number

Header

Pages
DISK-ORIENTED DBMS

Get page #2

Execution Engine

Memory

Buffer Pool

Disk

Database File

Directory

Header

1

2

3

4

5

Pages
DISK-ORIENTED DBMS

- Disk
- Memory
- Database File
- Buffer Pool

Get page #2
**DISK-ORIENTED DBMS**

Get page #2

Execution Engine

- Memory
- Buffer Pool
- Database File
- Pages
Disk-Oriented DBMS

Memory

Buffer Pool

Directory

Header

Page #2

Get page #2

Pointer to page #2

Execution Engine

Disk

Database File

Directory

Header

Page 1

Page 2

Page 3

Page 4

Page 5

Pages
**DISK-ORIENTED DBMS**

- **Memory**
  - Buffer Pool
    - Directory
    - Header
    - Page 2

- **Disk**
  - Database File
    - Directory
    - Header
    - Pages: 1, 2, 3, 4, 5, ...

- **Execution Engine**
  - Get page #2
  - Pointer to page #2
  - Interpret the layout of page #2...
DISK-ORIENTED DBMS

Memory
Buffer Pool

Execution Engine

Get page #2
Pointer to page #2
Interpret the layout of page #2...

Disk

Database File

Directory
Header

Page #2

Lectures 3-4

Pages

1 2 3 4 5 ...

Get page #2

Pointer to page #2

Interpret the layout of page #2...
DISK-ORIENTED DBMS

Memory

Disk

Buffer Pool

Directory

Header

Get page #2

Pointer to page #2

Interpret the layout of page #2...

Lecture 5

Execution Engine

Lectures 3-4

Pages

Directory

Header

1

2

3

4

5

...
DISK-ORIENTED DBMS

Memory

Buffer Pool

Lecture 5

Get page #2

Pointer to page #2

Interpret the layout of page #2...

Execution Engine

Disk

Database File

Directory

Header

Pages

Lectures 3-4

Lectures 11-12
WHY NOT USE THE OS?

The DBMS can use memory mapping (mmap) to store the contents of a file into the address space of a program.

The OS is responsible for moving the pages of the file in and out of memory, so the DBMS doesn’t need to worry about it.
WHY NOT USE THE OS?

The DBMS can use memory mapping (mmap) to store the contents of a file into the address space of a program.

The OS is responsible for moving the pages of the file in and out of memory, so the DBMS doesn’t need to worry about it.
WHY NOT USE THE OS?

The DBMS can use memory mapping (mmap) to store the contents of a file into the address space of a program.

The OS is responsible for moving the pages of the file in and out of memory, so the DBMS doesn’t need to worry about it.
WHY NOT USE THE OS?

The DBMS can use memory mapping (mmap) to store the contents of a file into the address space of a program.

The OS is responsible for moving the pages of the file in and out of memory, so the DBMS doesn’t need to worry about it.
WHY NOT USE THE OS?

The DBMS can use memory mapping (mmap) to store the contents of a file into the address space of a program. The OS is responsible for moving the pages of the file in and out of memory, so the DBMS doesn’t need to worry about it.
WHY NOT USE THE OS?

The DBMS can use memory mapping (mmap) to store the contents of a file into the address space of a program.

The OS is responsible for moving the pages of the file in and out of memory, so the DBMS doesn’t need to worry about it.
WHY NOT USE THE OS?

The DBMS can use memory mapping (mmap) to store the contents of a file into the address space of a program.

The OS is responsible for moving the pages of the file in and out of memory, so the DBMS doesn’t need to worry about it.
WHY NOT USE THE OS?

The DBMS can use memory mapping (mmap) to store the contents of a file into the address space of a program.

The OS is responsible for moving the pages of the file in and out of memory, so the DBMS doesn’t need to worry about it.
WHY NOT USE THE OS?

The DBMS can use memory mapping (mmap) to store the contents of a file into the address space of a program.

The OS is responsible for moving the pages of the file in and out of memory, so the DBMS doesn’t need to worry about it.
WHY NOT USE THE OS?

The DBMS can use memory mapping (mmap) to store the contents of a file into the address space of a program.

The OS is responsible for moving the pages of the file in and out of memory, so the DBMS doesn’t need to worry about it.
WHY NOT USE THE OS?

The DBMS can use memory mapping (mmap) to store the contents of a file into the address space of a program.

The OS is responsible for moving the pages of the file in and out of memory, so the DBMS doesn’t need to worry about it.
WHY NOT USE THE OS?

The DBMS can use memory mapping (mmap) to store the contents of a file into the address space of a program.

The OS is responsible for moving the pages of the file in and out of memory, so the DBMS doesn’t need to worry about it.
WHY NOT USE THE OS?

What if we allow multiple threads to access the `mmap` files to hide page fault stalls?

This works good enough for read-only access. It is complicated when there are multiple writers...
WHY NOT USE THE OS?

There are some solutions to this problem:

→ **madvise**: Tell the OS how you expect to read certain pages.

→ **mlock**: Tell the OS that memory ranges cannot be paged out.

→ **msync**: Tell the OS to flush memory ranges out to disk.
WHY NOT USE THE OS?

There are some solutions to this problem:

→ **madvise**: Tell the OS how you expect to read certain pages.

→ **mlock**: Tell the OS that memory ranges cannot be paged out.

→ **msync**: Tell the OS to flush memory ranges out to disk.
WHY NOT USE THE OS?

There are some solutions to this problem:

→ **madvise**: Tell the OS how you expect to read certain pages.

→ **mlock**: Tell the OS that memory ranges cannot be paged out.

→ **msync**: Tell the OS to flush memory ranges out to disk.
WHY NOT USE THE OS?

DBMS (almost) always wants to control things itself and can do a better job than the OS.

→ Flushing dirty pages to disk in the correct order.
→ Specialized prefetching.
→ Buffer replacement policy.
→ Thread/process scheduling.

The OS is not your friend.
Problem #1: How the DBMS represents the database in files on disk.

Problem #2: How the DBMS manages its memory and moves data back-and-forth from disk.
DATABASE STORAGE

Problem #1: How the DBMS represents the database in files on disk.

Problem #2: How the DBMS manages its memory and moves data back-and-forth from disk.

← Today
**DATABASE STORAGE**

**Problem #1:** How the DBMS represents the database in files on disk.

**Problem #2:** How the DBMS manages its memory and moves data back-and-forth from disk.
TODAY'S AGENDA

File Storage
Page Layout
Tuple Layout
FILE STORAGE

The DBMS stores a database as one or more files on disk typically in a proprietary format.
→ The OS doesn't know anything about the contents of these files.

Early systems in the 1980s used custom filesystems on raw storage.
→ Some "enterprise" DBMSs still support this.
→ Most newer DBMSs do not do this.
The storage manager is responsible for maintaining a database's files.
→ Some do their own scheduling for reads and writes to improve spatial and temporal locality of pages.

It organizes the files as a collection of pages.
→ Tracks data read/written to pages.
→ Tracks the available space.
A page is a fixed-size block of data.
→ It can contain tuples, meta-data, indexes, log records…
→ Most systems do not mix page types.
→ Some systems require a page to be self-contained.

Each page is given a unique identifier.
→ The DBMS uses an indirection layer to map page IDs to physical locations.
There are three different notions of "pages" in a DBMS:
→ Hardware Page (usually 4KB)
→ OS Page (usually 4KB)
→ Database Page (512B-16KB)

A hardware page is the largest block of data that the storage device can guarantee failsafe writes.
DATABASE PAGES

There are three different notions of "pages" in a DBMS:

→ Hardware Page (usually 4KB)
→ OS Page (usually 4KB)
→ Database Page (512B-16KB)

A hardware page is the largest block of data that the storage device can guarantee failsafe writes.
There are three different notions of "pages" in a DBMS:
→ Hardware Page (usually 4KB)
→ OS Page (usually 4KB)
→ Database Page (512B-16KB)

A hardware page is the largest block of data that the storage device can guarantee failsafe writes.
A heap file is an unordered collection of pages with tuples that are stored in random order.

→ Create / Get / Write / Delete Page
→ Must also support iterating over all pages.

Two ways to represent a heap file:
→ Linked List
→ Page Directory
DATABASE HEAP

It is easy to find pages if there is only a single heap file.

Need meta-data to keep track of what pages exist in multiple files and which ones have free space.
DATABASE HEAP

It is easy to find pages if there is only a single heap file.

Need meta-data to keep track of what pages exist in multiple files and which ones have free space.
It is easy to find pages if there is only a single heap file.

Need meta-data to keep track of what pages exist in multiple files and which ones have free space.
DATABASE HEAP

It is easy to find pages if there is only a single heap file.

Need meta-data to keep track of what pages exist in multiple files and which ones have free space.
DATABASE HEAP

It is easy to find pages if there is only a single heap file.

Need meta-data to keep track of what pages exist in multiple files and which ones have free space.
HEAP FILE: LINKED LIST

Maintain a header page at the beginning of the file that stores two pointers:
→ HEAD of the free page list.
→ HEAD of the data page list.

Each page keeps track of how many free slots they currently have.
Maintain a header page at the beginning of the file that stores two pointers:
→ HEAD of the free page list.
→ HEAD of the data page list.

Each page keeps track of how many free slots they currently have.
Maintain a **header page** at the beginning of the file that stores two pointers:

→ HEAD of the **free page list**.

→ HEAD of the **data page list**.

Each page keeps track of how many free slots they currently have.
Maintain a header page at the beginning of the file that stores two pointers:

→ HEAD of the free page list.
→ HEAD of the data page list.

Each page keeps track of how many free slots they currently have.
HEAP FILE: LINKED LIST

Maintain a header page at the beginning of the file that stores two pointers:
→ HEAD of the free page list.
→ HEAD of the data page list.

Each page keeps track of how many free slots they currently have.
The DBMS maintains special pages that tracks the location of data pages in the database files.

The directory also records the number of free slots per page.

Must make sure that the directory pages are in sync with the data pages.
The DBMS maintains special pages that tracks the location of data pages in the database files.

The directory also records the number of free slots per page.

Must make sure that the directory pages are in sync with the data pages.
The DBMS maintains special pages that tracks the location of data pages in the database files.

The directory also records the number of free slots per page.

Must make sure that the directory pages are in sync with the data pages.
TODAY'S AGENDA

File Storage
Page Layout
Tuple Layout
Every page contains a header of meta-data about the page's contents.

→ Page Size
→ Checksum
→ DBMS Version
→ Transaction Visibility
→ Compression Information

Some systems require pages to be self-contained (e.g., Oracle).
For any page storage architecture, we now need to decide how to organize the data inside of the page.
→ We are still assuming that we are only storing tuples.

Two approaches:
→ Tuple-oriented
→ Log-structured
For any page storage architecture, we now need to decide how to organize the data inside of the page. → We are still assuming that we are only storing tuples.

Two approaches:

→ Tuple-oriented
→ Log-structured
TUPLE STORAGE

How to store tuples in a page?
How to store tuples in a page?

**Strawman Idea:** Keep track of the number of tuples in a page and then just append a new tuple to the end.
TUPLE STORAGE

How to store tuples in a page?

**Strawman Idea:** Keep track of the number of tuples in a page and then just append a new tuple to the end.
How to store tuples in a page?

**Strawman Idea:** Keep track of the number of tuples in a page and then just append a new tuple to the end.

→ What happens if we delete a tuple?
## TUPLE STORAGE

How to store tuples in a page?

**Strawman Idea:** Keep track of the number of tuples in a page and then just append a new tuple to the end.

→ What happens if we delete a tuple?

<table>
<thead>
<tr>
<th>Page</th>
<th>Num Tuples = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tuple #1</td>
</tr>
<tr>
<td></td>
<td>Tuple #3</td>
</tr>
</tbody>
</table>
TUPLE STORAGE

How to store tuples in a page?

Strawman Idea: Keep track of the number of tuples in a page and then just append a new tuple to the end. → What happens if we delete a tuple?

<table>
<thead>
<tr>
<th>Num Tuples = 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuple #4</td>
</tr>
<tr>
<td>Tuple #3</td>
</tr>
<tr>
<td>Tuple #1</td>
</tr>
</tbody>
</table>
How to store tuples in a page?

**Strawman Idea:** Keep track of the number of tuples in a page and then just append a new tuple to the end.

→ What happens if we delete a tuple?
→ What happens if we have a variable-length attribute?
The most common layout scheme is called **slotted pages**.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:
- The # of used slots
- The offset of the starting location of the last slot used.
The most common layout scheme is called slotted pages.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:
→ The # of used slots
→ The offset of the starting location of the last slot used.
The most common layout scheme is called **slotted pages**.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:
- The # of used slots
- The offset of the starting location of the last slot used.
The most common layout scheme is called **slotted pages**.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:
→ The # of used slots
→ The offset of the starting location of the last slot used.
The most common layout scheme is called slotted pages.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:
→ The # of used slots
→ The offset of the starting location of the last slot used.
The most common layout scheme is called **slotted pages**.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:

- The # of used slots
- The offset of the starting location of the last slot used.
The most common layout scheme is called **slotted pages**.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:
→ The # of used slots
→ The offset of the starting location of the last slot used.
The most common layout scheme is called slotted pages.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:
→ The # of used slots
→ The offset of the starting location of the last slot used.
SLOTTED PAGES

The most common layout scheme is called *slotted pages*.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:

→ The # of used slots
→ The offset of the starting location of the last slot used.
The most common layout scheme is called slotted pages.

The slot array maps "slots" to the tuples' starting position offsets.

The header keeps track of:
→ The # of used slots
→ The offset of the starting location of the last slot used.
The DBMS needs a way to keep track of individual tuples. Each tuple is assigned a unique record identifier.

→ Most common: page_id + offset/slot

→ Can also contain file location info.

An application cannot rely on these IDs to mean anything.
RECORD IDS

The DBMS needs a way to keep track of individual tuples.

Each tuple is assigned a unique record identifier.
→ Most common: page_id + offset/slot
→ Can also contain file location info.

An application cannot rely on these IDs to mean anything.
TODAY'S AGENDA

File Storage
Page Layout
Tuple Layout
A tuple is essentially a sequence of bytes.

It's the job of the DBMS to interpret those bytes into attribute types and values.
Each tuple is prefixed with a header that contains meta-data about it.

- Visibility info (concurrency control)
- Bit Map for NULL values.

We do not need to store meta-data about the schema.
Attributes are typically stored in the order that you specify them when you create the table.

This is done for software engineering reasons (i.e., simplicity).

However, it might be more efficient to lay them out differently.
DENORMALIZED TUPLE DATA

DBMS can physically *denormalize* (e.g., "pre join") related tuples and store them together in the same page.

→ Potentially reduces the amount of I/O for common workload patterns.

→ Can make updates more expensive.

```
CREATE TABLE foo (  
a INT PRIMARY KEY,
  b INT NOT NULL,
);
```

```
CREATE TABLE bar (  
c INT PRIMARY KEY,
  a INT
REFERENCES foo (a),
);
```
DBMS can physically *denormalize* (e.g., "pre join") related tuples and store them together in the same page. → Potentially reduces the amount of I/O for common workload patterns. → Can make updates more expensive.

```sql
CREATE TABLE foo (  
a INT PRIMARY KEY,
  b INT NOT NULL,
);
CREATE TABLE bar (  
c INT PRIMARY KEY,
  a INT
REFERENCES foo (a),
);
```
DENORMALIZED TUPLE DATA

DBMS can physically *denormalize* (e.g., "pre join") related tuples and store them together in the same page.

→ Potentially reduces the amount of I/O for common workload patterns.

→ Can make updates more expensive.
DENORMALIZED TUPLE DATA

DBMS can physically *denormalize* (e.g., "pre join") related tuples and store them together in the same page.

→ Potentially reduces the amount of I/O for common workload patterns.

→ Can make updates more expensive.
DENORMALIZED TUPLE DATA

DBMS can physically *denormalize* (e.g., "pre join") related tuples and store them together in the same page.

→ Potentially reduces the amount of I/O for common workload patterns.
→ Can make updates more expensive.
DENORMALIZED TUPLE DATA

DBMS can physically **denormalize** (e.g., "pre join") related tuples and store them together in the same page.
→ Potentially reduces the amount of I/O for common workload patterns.
→ Can make updates more expensive.

Not a new idea.
→ IBM System R did this in the 1970s.
→ Several NoSQL DBMSs do this without calling it physical denormalization.
DENORMALIZED TUPLE DATA

DBMS can physically *denormalize* (e.g., "pre join") related tuples and store them together in the same page.

→ Potentially reduces the amount of I/O for common workload patterns.
→ Can make updates more expensive.

Not a new idea.
→ IBM System R did this in the 1970s.
→ Several NoSQL DBMSs do this without calling it physical denormalization.
CONCLUSION

Database is organized in pages.
Different ways to track pages.
Different ways to store pages.
Different ways to store tuples.
NEXT CLASS

Log-Structured Storage
Value Representation
Storage Models