COSC2406/2407: Database Systems

Disks and Files

Assignment Project Exam Help

Xiangmin (Emily) Zhou

Onli https://eduassistpro.github.i

Thursdays

Email : xiangmin.zhou@rm

Add WeChat edu_assist_pr

Lecture 2

References: Ramakrishnan & Gehrke Chapter 9
Garcia-Molina et al. Chapter 11
Elmasri & Navathe Chapters 5 & 6

Overview

Assignment Project Exam Help

Over the next two lectures, we will lay the foundations for studying database systems. We will discuss:

- Dis tics, and
- Op prin systems (DBMSs)
- Files—how files are allocated and manag
- Data and records—how data and records a

https://eduassistpro.github.i
Add WeChat edu_assist_pr

Representing Data Elements

First we discuss disks, their characteristics, and how DBMS buffer managers interact with disks.
We focus on blocks, and next week we focus on how data is stored in those blo

1. Attr byt
2. Fields are stored together to form logical *records*
3. Records are stored in disk blocks
4. Blocks of records of the same type are typical form a *file* (note that a file in a DBMS is differ operating system file)

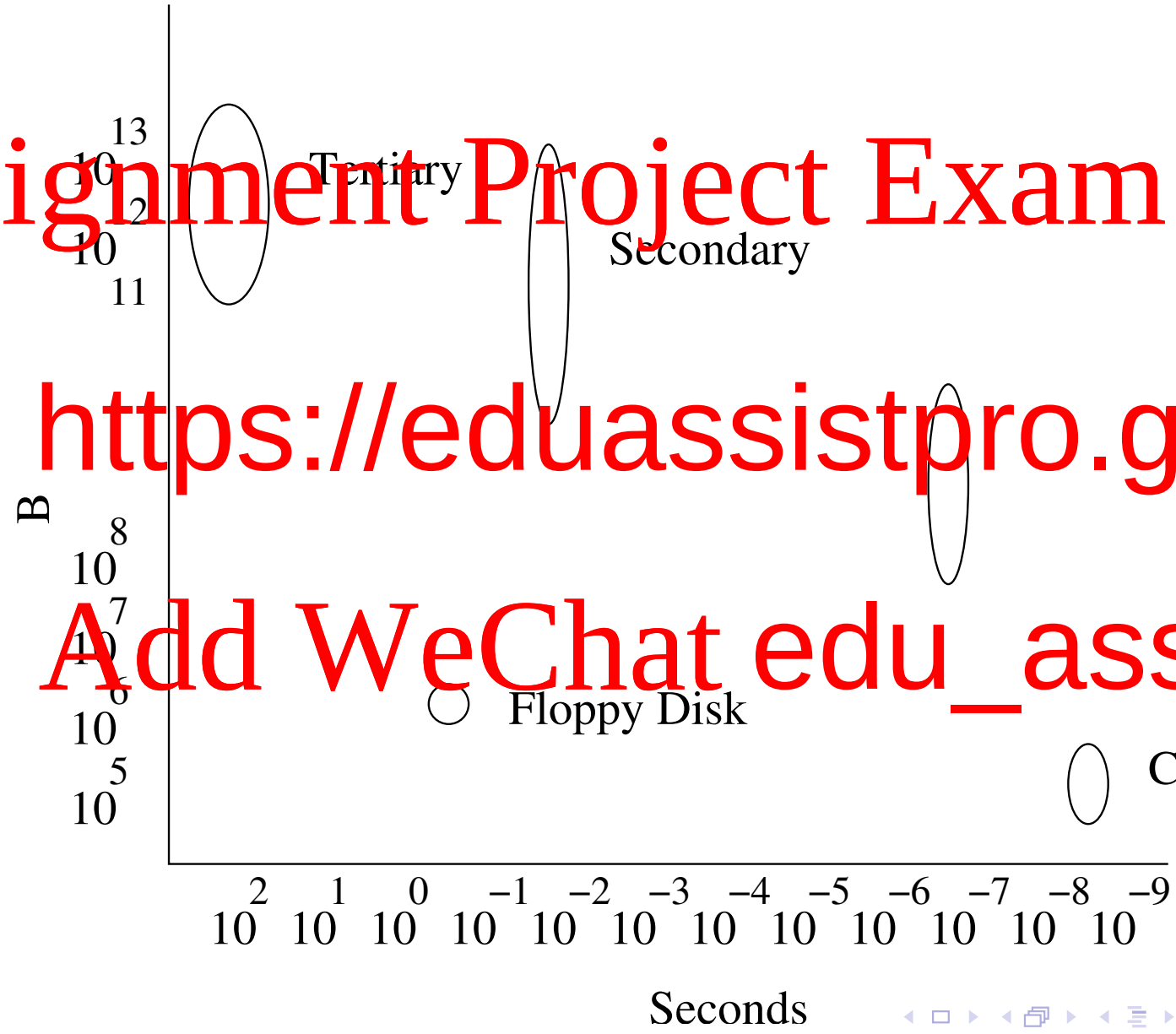
Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Memory Hierarchy

Cache is the lowest level of the hierarchy. Two components form the cache: on-board cache, on the same chip as the CPU, and level-2 cache on another chip. The typical maximum cache size is around one megabyte. A cache can be accessed in a few nanoseconds.
Main Memories are next in the hierarchy. A typical capacity is a few hundred s.
Main me
Second el
gigabyt
supports random access. Access takes aroun
and typical capacities are in tens of gigabytes.
Tertiary storage is cheaper still and slower ag
tapes—DLT, DAT, and so on—that are capable of very large storage capacities (perhaps terabytes). Access times are perhaps seconds or minutes.
Data is typically stored on disks and brought to main memory for processing by the Database Management System (DBMS).

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Memory Hierarchy



Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Secondary Storage - HDD versus SSD

Assignment Project Exam Help

- HDD = Hard Disk Drive
- SSD = Solid State Drive
- Ev
  De
  Go https://eduassistpro.github.i

  `https://www.youtube.com/watch?`
- HDD vs SSD - What is the difference?
  Carey Holzman (9 Feb 2015)

  `https://www.youtube.com/watch?`

Add WeChat edu_assist_pr

HDD: Characteristics

Disk storage *was* the most common non-volatile storage media for large amounts of data.

A single disk surface is divided into *tracks*, with each track containing as many a

surface.

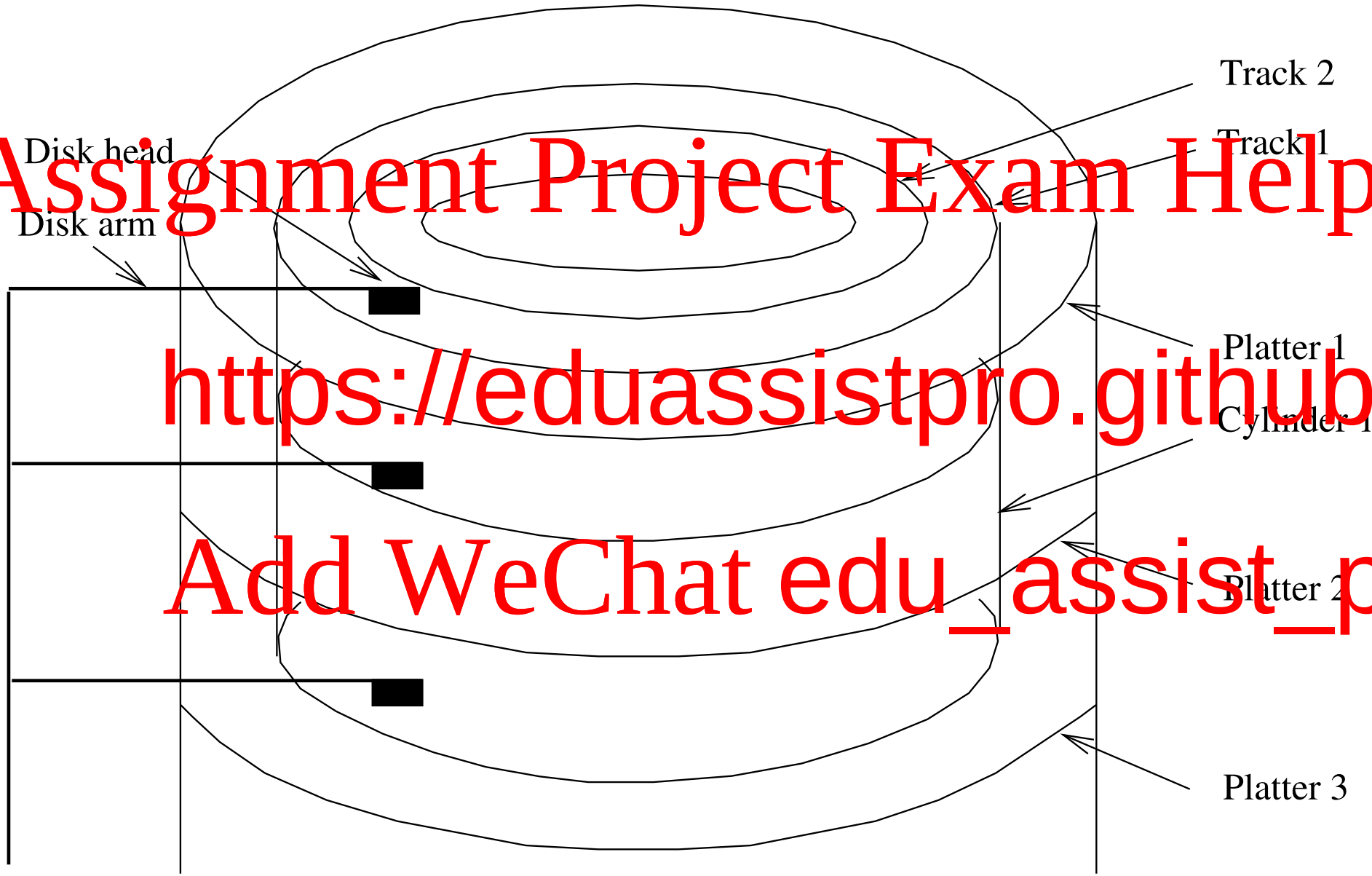
Multi-pl *ders*.

The time to find a track and set-up to read or write is known as the *seek time*, whilst the spin-time to find data is calle *ncy* (more in a moment).

Latency exists between tracks in a cylinder, as th aligned.

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

HDD: Sectors, Tracks, Platters and Cylinders



Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

HDD Disks

Assignment Project Exam Help

Data is stored in physical *sectors*.

A physical sector is a fixed length unit of storage that can be address

The sect

sector st

Logical

size is usually set when the disk is formatted or initia

(See Section 9.1.1 of Ramakrishnan & Gehrke fo

disks.)

https://eduassistpro.github.i

Add WeChat edu_assist_pr

HDD Performance

The access time for a block on disk has 3 main components:

- seek time
- rotational delay
- transfer time

Assignment Project Exam Help

https://eduassistpro.github.i

Add WeChat edu_assist_pr

For typical HDD, the *block access time* is 10 to 15 milliseconds. Since seek time and rotational delays dominate the total, the time to read one block of data is almost the same as that of reading several *contiguous* blocks. This is sometimes referred to as *blocked access*.

HDD Performance

Assignment Project Exam Help

Consider a simple disk that has a 10 ms seek time, 8 kb blocks, and can read 1

A: How lo

https://eduassistpro.github.i

B: How lo

C: How long will it take to read ten non-contiguous blocks?

What does this suggest about record organisati

Add WeChat edu_assist_pr

HDD Performance

Assignment Project Exam Help

The simple disk has a 10 ms seek time, 8 kb blocks, and can read 10 Mb per second from disk.

A: Readi

$0.01 + \frac{1}{}$

https://eduassistpro.github.i

Ten conti

$0.01 + 10 \times 0.78 = 17.8$ ms

Ten non-contiguous blocks: Ten seeks, ten rea

$10 \times 0.1 + 10 \times 0.78 = 167.8$ ms

Add WeChat edu_assist_pr

HDD Performance

Assignment Project Exam Help

Consider a disk that rotates at 7,200 rpm (it makes a rotation in 60/7200 = 0.00833 seconds or 8.33 ms). The block size is 16,384 bytes, an
sectors p
The head
for every 1
track in 2.001 ms, or from the innermost track to the o
$2 + 16384/1000 = 18.384$ ms.
What are the approximate minimum and maxim
block?

https://eduassistpro.github.i

Add WeChat edu_assist_pr

HDD Performance

Assignment Project Exam Help

Minimum time: the time to read when the head is positioned to read the required block. So, 4 sectors of the 128 sectors require $4/128 \times$

Maximu

block is as f

to move th e

just missed the start of the desired block). Finally, $4/128 \times 8.33 = 0.26$ ms to read. Total: $18.38 + 8.33 + 0.26 = 26.97$ ms.

The average time is harder: see Example 11.5 in th

https://eduassistpro.github.i

Add WeChat edu_assist_pr

Improving HDD performance

The Elevator Algorithm

In practice, disks have a queue of requests for blocks.

One approach is to process them in order.

Better approach is the *elevator algorithm*, so-called because it schedules block accesses as the disk arm sweeps back and forth across th

- Wh
  req
- Wh
  sweeps in the opposite direction.

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Organising Data by Cylinders

Since seek time represents about half average ti
it makes sense to store data that is likely to be accessed together (such as a relation in a database), in a single cylinder.

If there is not enough room, then use several adjacent cylinders.

If reading whole cylinder, only need one seek (to move to the cylinder) and first rotational latency (until first block moves under the head).

Improving HDD performance: Striping Disks

Assignment Project Exam Help

Striping creates a single logical volume from two or more disks. As an example, when tracks are striped on two disks, odd numbered tracks come from one disk and even numbered tracks from the other.

The princ

improve

accesse

https://eduassistpro.github.i

disks. However, this can affect other disk activities.

Striping can be in units of tracks, blocks, cylinders so on (see section 9.2.1).

Add WeChat edu_assist_pr

(Striping without redundant storage is so-calle

RAID on the next slide.)

Improving redundancy and performance: Disk Arrays

RAIDs (*Redundant Arrays of Inexpensive Disks*) are a set of disk drives that are accessed concurrently to increase transfer rates and the number of possible concurrent accesses.
RAID also offers other features in so-called RAID-1 through to RAID-6.
Differen
perform
two identi
Solomo
(you should read and understand section 9.2.3)

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

RAID provides both redundancy and performa

RAID also can be and is used to provide redundancy for SSD, but as not required for performance with SSD the technology is evolving.

RAID-3

Assignment Project Exam Help

For a set of n disks operating under RAID-3, one disk is set aside to store only the parity of the other $n-1$ disks. For example, consider the first bit on e
i of the dis
if i is odd
corresp

https://eduassistpro.github.i

If any disk fails, the parity disk and other disks can be u
reconstruct the failed disk: the bit in any position is th
of the the bits in the corresponding position of the s.

Add WeChat edu_assist_pr

Improving HDD performance: Disk Caches

Disks (and file systems) are typically designed to reduce the number of disk accesses required to retrieve data; memory accesses are much cheaper than disk accesses.

A *disk cache* is a fast storage device that keeps data in memory from the last *n* disk accesses.

All I/O tran
cache an
General on
on Least Recently Used (LRU): every reference to a block in the cache moves that block to the end of the "replacement qu
A block read will succeed without a disk access if th
cache; just needs a simple memory copy from cac

Disk caches are capable of detecting whether recent accesses were sequential; if so, pre-fetch data blocks in anticipation.

Because of the delayed writing and anticipated reading scheme, such cache algorithms are known as *read-ahead, write-behind* caching.

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

HDD Performance Improvement

Assignment Project Exam Help

In summary, several techniques are used to improve disk access performance:

- Org
- Usi
- Stri
- Using the elevator algorithm
- Pre-fetching data into a cache

https://eduassistpro.github.i
Add WeChat edu_assist_pr

Disk Performance in Practice

Disk characteristics in a real system.

Assignment Project Exam Help

```
iostat -x
                                  extended disk statistics
disk        r/s   w/s    Kr/s    Kw/s wait a
sd0         1.0   1.0     8.0    11.8  0.0  0.0   23.9   0   2
sd2         1.0   0.5     7.2     3.7  0.0  0.1   47.2   0   1
sd3         1.5   0.7    10.6     5.4  0.0  0.0   22.6   0   2
sd6         0.0   0.0     0.0     0.0  0.0  0.0  271.6   0   0
sd17        0.1   0.0     0.4     0.2  0.0  0.0   42.0   0   0
```

https://eduassistpro.github.i

Add WeChat edu_assist_pr

`svc_t` is the interesting column: it indicates how long, on average, the disk takes to respond to a request in milliseconds.

Disk Space Management in a DBMS

Assignment Project Exam Help

The lowest level of the DBMS software – the *disk space manager* – hides details of the disk hardware from higher levels of the DBMS softwar

pages (

Higher le

read or wri

The page size is set to be the same as the disk block siz

page is stored in a disk block, and a page read/write c

one disk I/O.

The terms “page” and “block” are sometimes use

https://eduassistpro.github.i

Add WeChat edu_assist_pr

Buffer Management in a DBMS

The *buffer manager*, a level above the disk space manager, ensures that pages requested by higher levels are present in main memory (see section 9.4 of the text).
The buffer manager partitions available main memory into *frames*.
One pag
buffer po
For each f
maintains a *pin-count* of the number of users
frame.
If a new process starts using a frame, its pin-count is i
pinning). When a process is finished with the fram
decremented. A page can only be swapped out of memory when the pin-count is zero.

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Buffer Management...

Assignment Project Exam Help

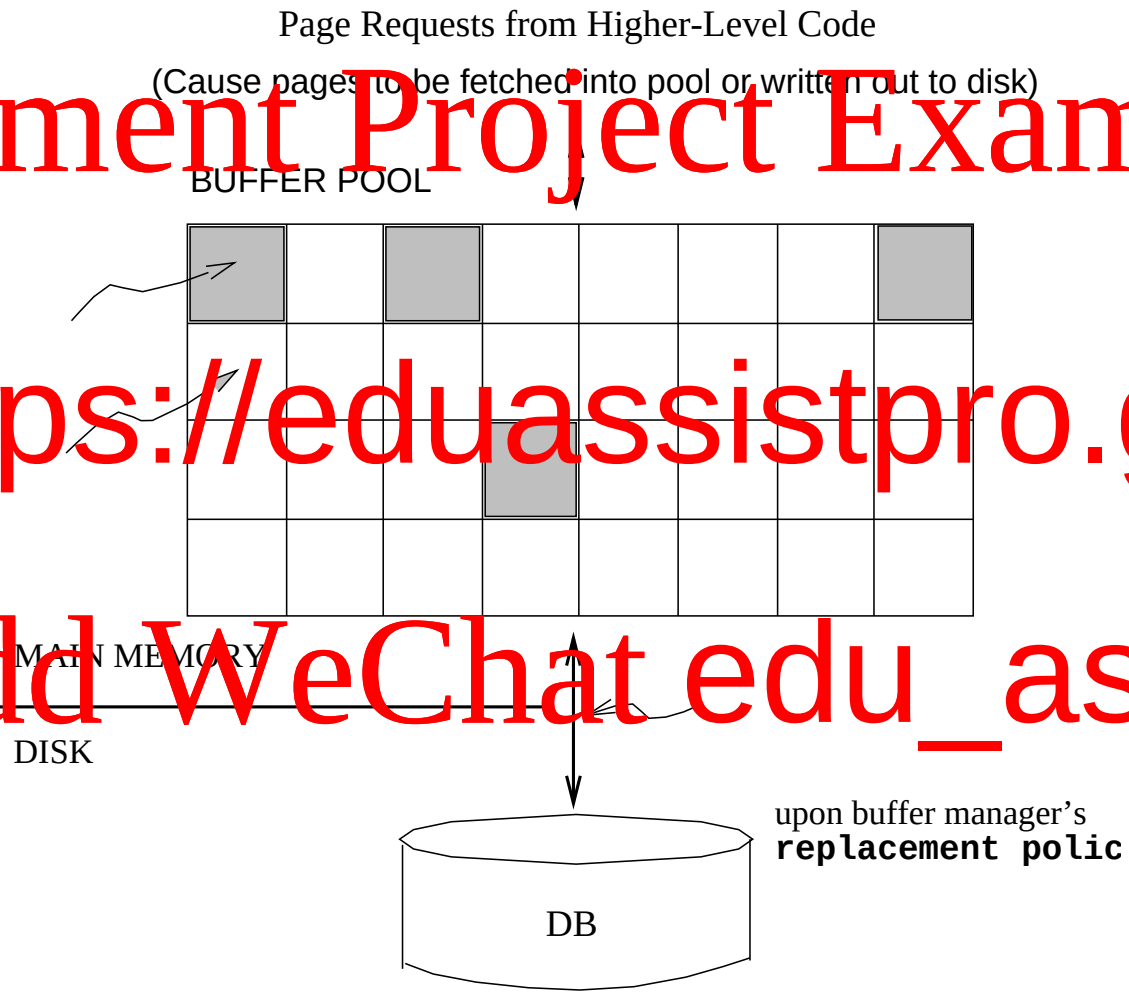
If the requestor has modified a page, it sets a "dirty" bit for that frame
Before the buffer manager can swap a page with a dirty bit out of memory

https://eduassistpro.github.i

A variety o
should b
(LRU), M
and Clock.
Different replacement policies are optimal in diff
(Read section 9.4.1 about the replacement algo

Add WeChat edu_assist_pr

Buffer Management in a DBMS



Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

DBMS Buffer Management in Practice

Assignment Project Exam Help

- Named pools (Sybase/DB2) – Pages can be grouped in pools, and different replacement policies assigned to the pools
- Diff nts, “ha
- Pre d can

https://eduassistpro.github.i

Oracle 8 uses prefetch for sequential scan does), retrieving large objects, and certain e on indexes later)

Add WeChat edu_assist_pr

Structured File Management in a DBMS

Assignment Project Exam Help

Low-level file management serves as a way of abstracting device drivers, h
useful, p
In a DBMS
function
DBMS file management aims to arrange and for
such that space and access costs are minimised

https://eduassistpro.github.i
Add WeChat edu_assist_pr

DBMS Disk Space Management

Assignment Project Exam Help

There are two main approaches to DBMS disk space management.

1. The DBMS depends on the OS file system. The whole DBMS is allo
spa
2. The
disks and memory itself

https://eduassistpro.github.i

In either case, one of the tasks is to keep track of free bl
done by bit maps or linked lists.

Add WeChat edu_assist_pr

DBMS vs. OS File system

If the OS does disk space and buffer management, why not let the OS manage these tasks for the DBMS?

- Diff
- So
- Buf
  -
  - Adjust the replacement policy, and pre-f
    access patterns in typical DB operations (*ence patterns*)

We discuss DBMS file organisation in the next lecture.

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Databases and Files

A database may contain many files.

A file itself is stored as a set of blocks on the disk, that is, the data is made up of many logical blocks.

How do w
allocate b

- By allocating contiguous blocks, and kee
  and total number of blocks
- By using a linked list
- By using a directory of blocks. Note that we could allocate clusters of contiguous blocks also in this case.

We discuss these in detail next.

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Unordered (Heap) Files

Assignment Project Exam Help

- Simplest structure contains *records* in no particular order (more on records later and file structures in the next lecture)
- As th
de-
- To s
  - keep track of blocks in a file
  - keep track of free space on blocks
  - keep track of *records* on blocks

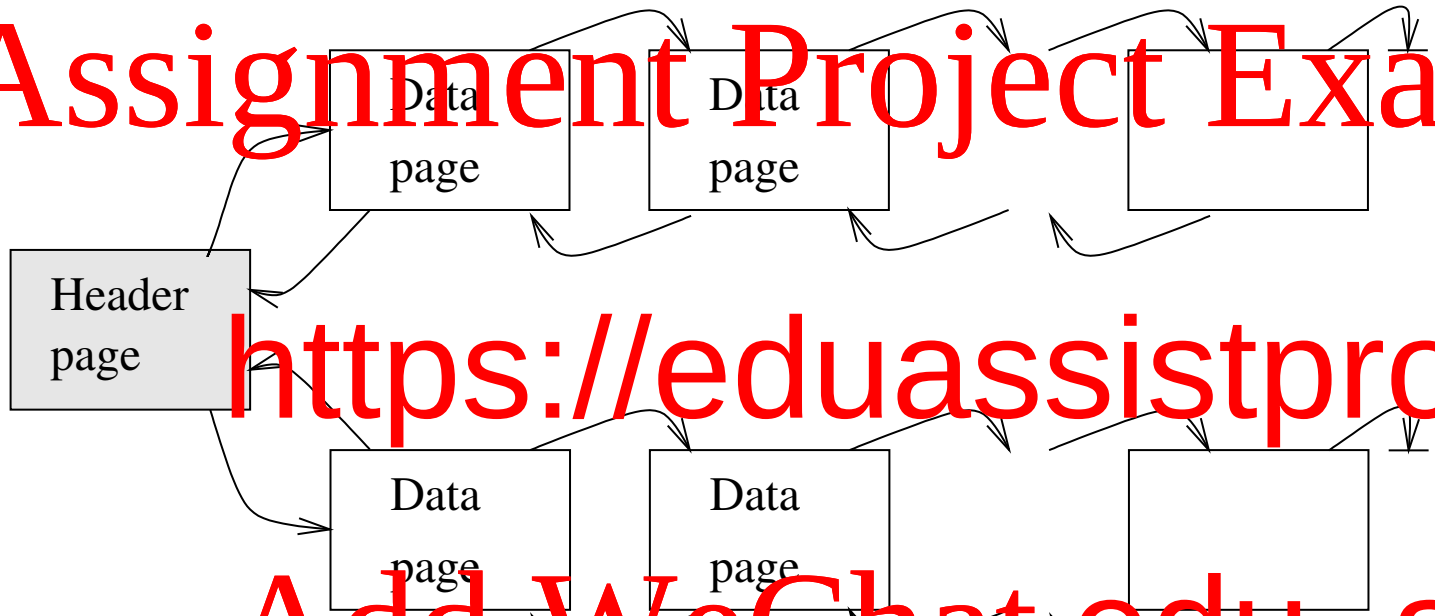
https://eduassistpro.github.i

Add WeChat edu_assist_pr

(There are many alternative solutions to these pr

Heap File Linked List Approach

Assignment Project Exam Help



https://eduassistpro.github.i

Add WeChat edu_assist_pr

One alternative is the heap file approach. Here, w
list of blocks that have free space—those where data can be inserted—and a second list of blocks that are full.

Heap File Linked List Approach...

If new blocks are required, a request is made of the buffer manager, and the new block is then added to the list of blocks in the file (probably at first as a block with free space).

Deletion:

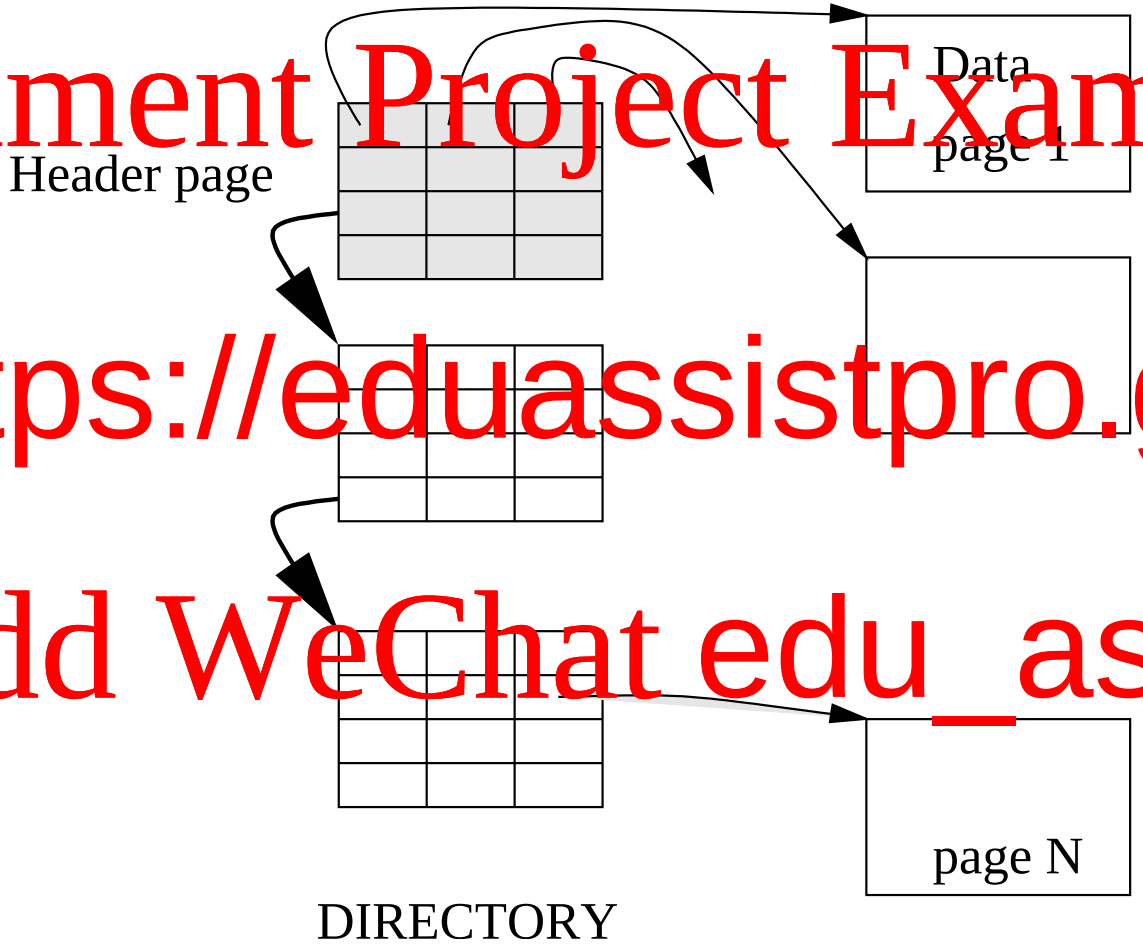
A disadvantage of this approach is that—because of internal fragmentation of the blocks—almost all blocks blocks with free space.

(The directory-based heap file solves the problem.)

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Heap File Block Directory Approach

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr



Heap File Block Directory Approach...

Assignment Project Exam Help

- The directory entry for a block can include the number of free bytes on the block
- The
imphttps://eduassistpro.github.i
- The
blocks can be allocated and deallocated as
- Directories contain little information and a small compared to the data blocks

Add WeChat edu_assist_pr

Summary: Disks and Files

Assignment Project Exam Help

In this lecture we have discussed:

- Me
- Dis
- Buf
- Buf
- Heap file organisation

https://eduassistpro.github.i

Next lecture, we will cover data records, and bag

Add WeChat edu_assist_pr