




CS 525: Advanced Database Organization

03: Disk Organization





Boris Glavic

Slides: adapted from a [course](#) taught by [Hector Garcia-Molina](#), Stanford InfoLab

CS 525

Notes 3
1




Topics for today

- How to lay out data on disk
- How to move it to/from memory

CS 525

Notes 3
2


What are the data items we want to store?


- a salary
- a name
- a date
- a picture



CS 525

Notes 3
3


What are the data items we want to store?

- a salary
- a name
- a date
- a picture

⇒ What we have available: Bytes



CS 525

Notes 3
4


To represent:

- Integer (short): 2 bytes
e.g., 35 is



00000000

00100011

Endian! Could as well be

00100011

00000000
- Real, floating point
 n bits for mantissa, m for exponent....



CS 525

Notes 3
5


To represent:

- Characters
→ various coding schemes suggested,
most popular is ASCII (1 byte encoding)

Example:

A: 1000001
a: 1100001
5: 0110101
LF: 0001010

CS 525

Notes 3
6


To represent:

- Boolean
e.g., TRUE

1111	1111
------	------

FALSE

0000	0000
------	------
- Application specific
e.g., enumeration
RED → 1 GREEN → 3
BLUE → 2 YELLOW → 4 ...

CS 525



Notes 3

7

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

To represent:

- Boolean
e.g., TRUE

1111	1111
------	------

FALSE

0000	0000
------	------
- Application specific
e.g., RED → 1 GREEN → 3
BLUE → 2 YELLOW → 4 ...

⇒ Can we use less than 1 byte/code?

Yes, but only if desperate...

CS 525



Notes 3

8

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

To represent:

- Dates
e.g.: - Integer, # days since Jan 1, 1900
- 8 characters, YYYYMMDD
- 7 characters, YYYYDDD
(not YYYYMMDD! Why?)
- Time
e.g. - Integer, seconds since midnight
- characters, HHMMSSFF

CS 525



Notes 3

9

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

To represent:

- String of characters
- Null terminated
e.g.,

c	a	t	⊗		
---	---	---	---	--	--
- Length given
e.g.,

3	c	a	t	⊗	
---	---	---	---	---	--
- Fixed length

CS 525



Notes 3

10

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

To represent:

- Bag of bits

Length	Bits
--------	------

CS 525



Notes 3

11

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Key Point

- Fixed length items
- Variable length items
- usually length given at beginning

CS 525




Notes 3

12

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY


Also

- Type of an item: Tells us how to interpret (plus size if fixed)

CS 525  Notes 3 13 IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY


Overview

Data Items
↓
Records
↓
Blocks
↓
Files
⋮
Memory

CS 525  Notes 3 14 IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY


Record - Collection of related data items (called FIELDS)

E.g.: Employee record:
name field,
salary field,
date-of-hire field, ...

CS 525  Notes 3 15 IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Types of records:


- Main choices:
 - FIXED vs VARIABLE FORMAT
 - FIXED vs VARIABLE LENGTH

CS 525  Notes 3 16 IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Fixed format

A SCHEMA (not record) contains following information

- # fields
- type of each field
- order in record
- meaning of each field

CS 525  Notes 3 17 IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Example: fixed format and length


Employee record

- (1) E#, 2 byte integer
- (2) E.name, 10 char.
- (3) Dept, 2 byte code

} Schema

55 | s m i t h | 02
83 | j o n e s | 01

} Records

CS 525  Notes 3 18 IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Variable format

- Record itself contains format
“Self Describing”

CS 525

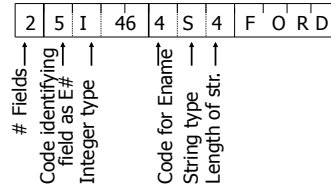


Notes 3

19

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Example: variable format and length



Field name codes could also be strings, i.e. TAGS

CS 525



Notes 3

20

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Variable format useful for:

- “sparse” records
- repeating fields
- evolving formats

.....→ But may waste space...
Additional indirection...

CS 525



Notes 3

21

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

- EXAMPLE:** var format record with repeating fields
Employee → one or more → children

3	E_name: Fred	Child: Sally	Child: Tom
---	--------------	--------------	------------

CS 525



Notes 3

22

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Note: Repeating fields does not imply

- variable format, nor
- variable size

John	Sailing	Chess	--
------	---------	-------	----

CS 525



Notes 3

23

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Note: Repeating fields does not imply

- variable format, nor
- variable size

John	Sailing	Chess	--
------	---------	-------	----

- Key is to allocate maximum number of repeating fields (if not used → null)

CS 525



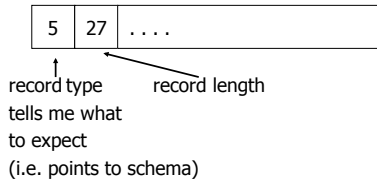
Notes 3

24

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

☆ Many variants between fixed - variable format:

Example: Include record type in record



Record header - data at beginning that describes record

May contain:

- record type
- record length
- time stamp
- null-value bitmap
- other stuff ...

Other interesting issues:

- Compression
 - within record - e.g. code selection
 - collection of records - e.g. find common patterns
- Encryption
- Splitting of large records
 - E.g., image field, store pointer

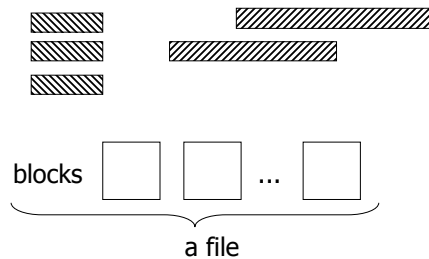
Record Header – null-map

- SQL: NULL is special value for every data type
 - Reserve one value for each data type as NULL?
- Easier solution
 - Record header has a bitmap to store whether field is NULL
 - Only store non-NULL fields in record

Separate Storage of Large Values

- Store fields with large values separately
 - E.g., image or binary document
 - Records have pointers to large field content
- Rationale
 - Large fields mostly not used in search conditions
 - Benefit from smaller records

Next: placing records into blocks



Next: placing records into blocks

assume fixed length blocks

blocks

a file — assume a single file (for now)

CS 525 COMPUTER SCIENCE Notes 3 31 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY

Options for storing records in blocks:

- (1) separating records
- (2) spanned vs. unspanned
- (3) sequencing
- (4) indirection

CS 525 COMPUTER SCIENCE Notes 3 32 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY

(1) Separating records

Block

- (a) no need to separate - fixed size recs.
- (b) special marker
- (c) give record lengths (or offsets)
 - within each record
 - in block header

CS 525 COMPUTER SCIENCE Notes 3 33 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY

(2) Spanned vs. Unspanned

- Unspanned: records must be within one block

- Spanned

CS 525 COMPUTER SCIENCE Notes 3 34 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY

With spanned records:

need indication of partial record "pointer" to rest

need indication of continuation (+ from where?)

CS 525 COMPUTER SCIENCE Notes 3 35 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY

Spanned vs. unspanned:

- Unspanned is much simpler, but may waste space...
- Spanned essential if record size > block size

CS 525 COMPUTER SCIENCE Notes 3 36 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY

(3) Sequencing

- Ordering records in file (and block) by some key value

Sequential file (\Rightarrow sequenced)

CS 525



Notes 3

37

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Why sequencing?

Typically to make it possible to efficiently read records in order
(e.g., to do a merge-join — discussed later)

CS 525



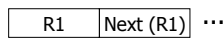
Notes 3

38

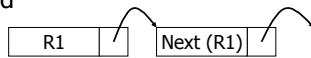
IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Sequencing Options

(a) Next record physically contiguous



(b) Linked



CS 525



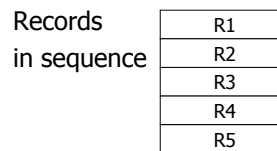
Notes 3

39

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Sequencing Options

(c) Overflow area



CS 525



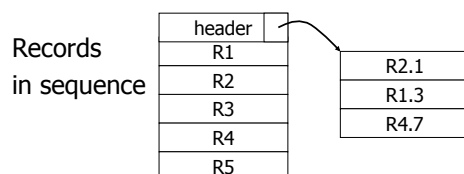
Notes 3

40

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Sequencing Options

(c) Overflow area



CS 525



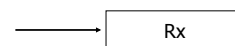
Notes 3

41

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

(4) Indirection

- How does one refer to records?



CS 525



Notes 3

42

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

(4) Indirection

- How does one refer to records?



Many options:

Physical \longleftrightarrow Indirect

CS 525



Notes 3

43

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

☆ Purely Physical

E.g., Record Address or ID = $\left\{ \begin{array}{l} \text{Device ID} \\ \text{Cylinder \#} \\ \text{Track \#} \\ \text{Block \#} \\ \text{Offset in block} \end{array} \right\}$ Block ID

CS 525



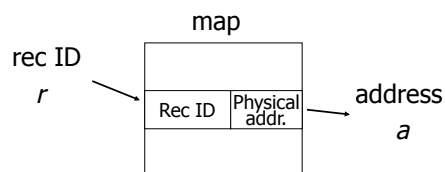
Notes 3

44

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

☆ Fully Indirect

E.g., Record ID is arbitrary bit string



CS 525



Notes 3

45

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Tradeoff

Flexibility \longleftrightarrow Cost
to move records of indirection
(for deletions, insertions)

CS 525



Notes 3

46

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Physical \longleftrightarrow Indirect

↑
Many options
in between ...

CS 525



Notes 3

47

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Block header - data at beginning that describes block

May contain:

- File ID (or RELATION or DB ID)
- This block ID
- Record directory
- Pointer to free space
- Type of block (e.g. contains recs type 4; is overflow, ...)
- Pointer to other blocks "like it"
- Timestamp ...

CS 525



Notes 3

48

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Example: Indirection in block

A block: Header

Free space

R4

R3

R1

R2

CS 525 COMPUTER SCIENCE Notes 3 49 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY

Tuple Identifier (TID)

- TID is
 - Page identifier
 - Slot number
- Slot stores either record or pointer (TID)
- TID of a record is fixed for all time

CS 525 COMPUTER SCIENCE Notes 3 50 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY

TID Operations

- Insertion
 - Set TID to record location (page, slot)
- Moving record
 - e.g., update variable-size or reorganization
 - Case 1: TID points to record
 - Replace record with pointer (new TID)
 - Case 2: TID points to pointer (TID)
 - Replace pointer with new pointer

CS 525 COMPUTER SCIENCE Notes 3 51 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY

TID: Block 1, Slot 2

Block 1

Block 2

CS 525 COMPUTER SCIENCE Notes 3 52 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY

Move record to Block 2 slot 3 -> TID does not change!

TID: Block 1, Slot 2

Block 1

Block 2

Block 2, Slot 3

CS 525 COMPUTER SCIENCE Notes 3 53 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY

Move record again to Block 2 slot 2 -> still one level of indirection

TID: Block 1, Slot 2

Block 1

Block 2

Block 2, Slot 2

CS 525 COMPUTER SCIENCE Notes 3 54 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY

TID Properties

- TID of record never changes
 - Can be used safely as pointer to record (e.g., in index)
- At most one level of indirection
 - Relatively efficient
 - Changes to physical address - changing max 2 pages

CS 525



Notes 3

55

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Options for storing records in blocks:

- (1) separating records
- (2) spanned vs. unspanned
- (3) sequencing
- (4) indirection

CS 525



Notes 3

56

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Other Topics

- (1) Insertion/Deletion
- (2) Buffer Management
- (3) Comparison of Schemes

CS 525



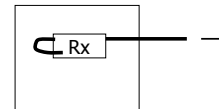
Notes 3

57

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Deletion

Block



CS 525



Notes 3

58

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Options:

- Immediately reclaim space
- Mark deleted

CS 525



Notes 3

59

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Options:

- Immediately reclaim space
- Mark deleted
 - May need chain of deleted records (for re-use)
 - Need a way to mark:
 - special characters
 - delete field
 - in map

CS 525



Notes 3

60

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

☆ As usual, many tradeoffs...

- How expensive is it to move valid record to free space for immediate reclaim?
- How much space is wasted?
 - e.g., deleted records, delete fields, free space chains,...

CS 525



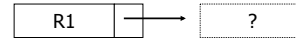
Notes 3

61

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Concern with deletions

Dangling pointers



CS 525



Notes 3

62

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Solution #1: Do not worry

CS 525



Notes 3

63

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Solution #2: Tombstones

E.g., Leave “MARK” in map or old location

CS 525



Notes 3

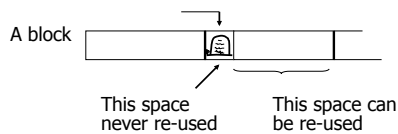
64

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Solution #2: Tombstones

E.g., Leave “MARK” in map or old location

- Physical IDs



CS 525



Notes 3

65

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Solution #2: Tombstones

E.g., Leave “MARK” in map or old location

- Logical IDs

map

ID	LOC
7788	

Never reuse ID 7788 nor space in map...

CS 525



Notes 3

66

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Insert

Easy case: records not in sequence

- Insert new record at end of file or in deleted slot
- If records are variable size, not as easy...

CS 525



Notes 3

67

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Insert

Hard case: records in sequence

- If free space “close by”, not too bad...
- Or use overflow idea...

CS 525



Notes 3

68

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Interesting problems:

- How much free space to leave in each block, track, cylinder?
- How often do I reorganize file + overflow?

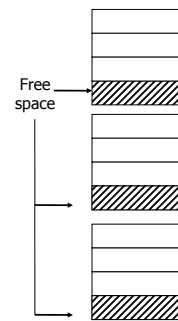
CS 525



Notes 3

69

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY



CS 525



Notes 3

70

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Buffer Management

- For Caching of Disk Blocks
- Buffer Replacement Strategies
 - E.g., LRU, clock
- Pinned blocks
- Forced output -----→ in Notes02
- Double buffering
- Swizzling

CS 525



Notes 3

71

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Buffer Manager

- Manages blocks cached from disk in main memory
- Usually -> fixed size buffer (M pages)
- DB requests page from Buffer Manager
 - Case 1: page is in memory -> return address
 - Case 2: page is on disk -> load into memory, return address

CS 525



Notes 3

72

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Goals

- Reduce the amount of I/O
- Maximize the *hit rate*
 - Ratio of number of page accesses that are fulfilled without reading from disk
- -> Need strategy to decide when to

CS 525



Notes 3

73

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Buffer Manager Organization

- Bookkeeping
 - Need to map (hash table) page-ids to locations in buffer (**page frames**)
 - Per page store *fix count, dirty bit, ...*
 - Manage free space
- Replacement strategy
 - If page is requested but buffer is full
 - Which page to emit remove from buffer

CS 525



Notes 3

74

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

FIFO

- **First In, First Out**
- Replace page that has been in the buffer for the longest time
- Implementation: E.g., pointer to oldest page (circular buffer)
 - $\text{Pointer} \rightarrow \text{next} = \text{Pointer}++ \% M$
- Simple, but not prioritizing frequently accessed pages

CS 525



Notes 3

75

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

LRU

- Least Recently Used
- Replace page that has not been accessed for the longest time
- Implementation:
 - List, ordered by LRU
 - Access a page, move it to list tail
- Widely applied and reasonable performance

CS 525



Notes 3

76

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Clock

- Frames are organized clock-wise
- Pointer *S* to current frame
- Each frame has a reference bit
 - Page is loaded or accessed -> bit = 1
- Find page to replace (advance pointer)
 - Return first frame with bit = 0
 - On the way set all bits to 0

CS 525

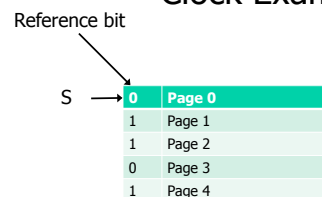


Notes 3

77

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Clock Example



CS 525



Notes 3

78

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Other Replacement Strategies

- LRU-K
- GCLOCK
- Clock-Pro
- ARC
- LFU

CS 525

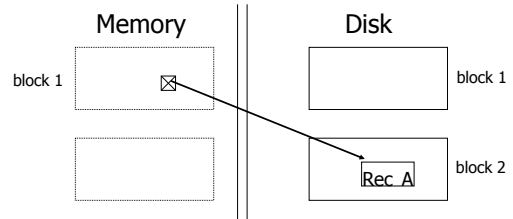


Notes 3

79

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Swizzling



CS 525

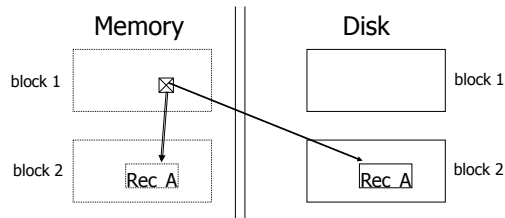


Notes 3

80

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Swizzling



CS 525



Notes 3

81

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Row vs Column Store

- So far we assumed that fields of a record are stored contiguously (row store)...
- Another option is to store all values of a field together (column store)

CS 525



Notes 3

82

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Row Store

- Example: Order consists of
 - id, cust, prod, store, price, date, qty

id1	cust1	prod1	store1	price1	date1	qty1
id2	cust2	prod2	store2	price2	date2	qty2
id3	cust3	prod3	store3	price3	date3	qty3

CS 525



Notes 3

83

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Column Store

- Example: Order consists of
 - id, cust, prod, store, price, date, qty

id1	cust1	id1	prod1	id1	price1	qty1
id2	cust2	id2	prod2	id2	price2	qty2
id3	cust3	id3	prod3	id3	price3	qty3
id4	cust4	id4	prod4	id4	price4	qty4
...

ids may or may not be stored explicitly

CS 525



Notes 3

84

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Row vs Column Store

- Advantages of Column Store
 - more compact storage (fields need not start at byte boundaries)
 - Efficient compression, e.g., RLE
 - efficient reads on data mining operations
- Advantages of Row Store
 - writes (multiple fields of one record) more efficient
 - efficient reads for record access (OLTP)

CS 525



Notes 3

85

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Comparison

- There are 10,000,000 ways to organize my data on disk...

Which is right for me?

CS 525

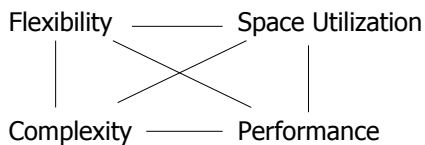


Notes 3

86

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Issues:



CS 525



Notes 3

87

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

- ☆ To evaluate a given strategy, compute following parameters:

- > space used for expected data
- > expected time to

- fetch record given key
- fetch record with next key
- insert record
- append record
- delete record
- update record
- read complete file
- reorganize file

CS 525



Notes 3

88

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Example

How would you design Megatron 3000 storage system? (for a relational DB, low end)

- Variable length records?
- Spanned?
- What data types?
- Fixed format?
- Record IDs ?
- Sequencing?
- How to handle deletions?

CS 525



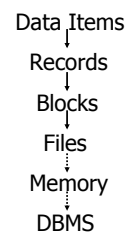
Notes 3

89

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Summary

- How to lay out data on disk



CS 525



Notes 3

90

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Next

How to find a record quickly,
given a key

CS 525



Notes 3

91

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY