# 01418231
# Data Structure

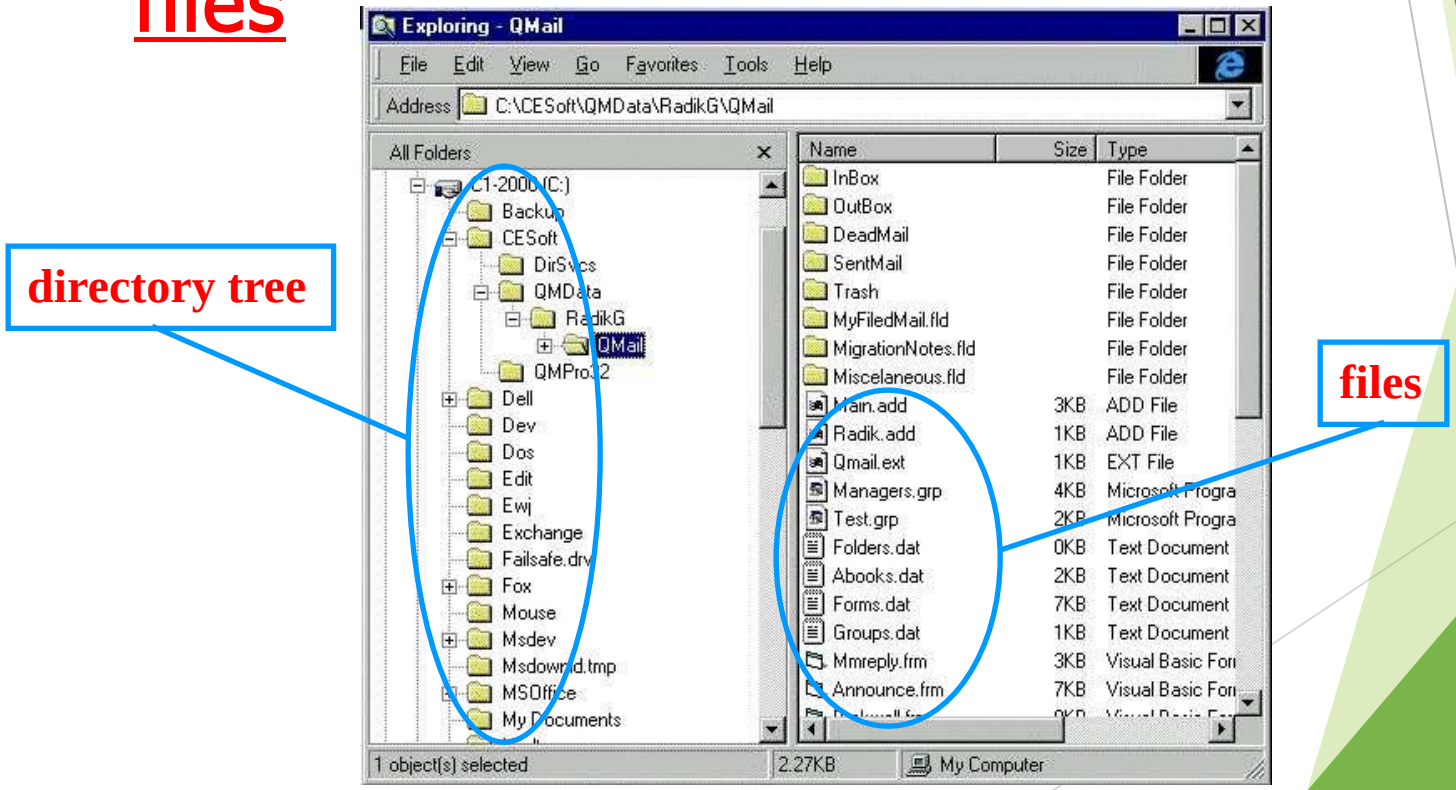
## File

# Agenda

- Overview of Files
- Field and record organization
- Type of Data Processing
- Access  Method
  - Sequential file
  - Random file
    - Indexed files
    - Hashed files

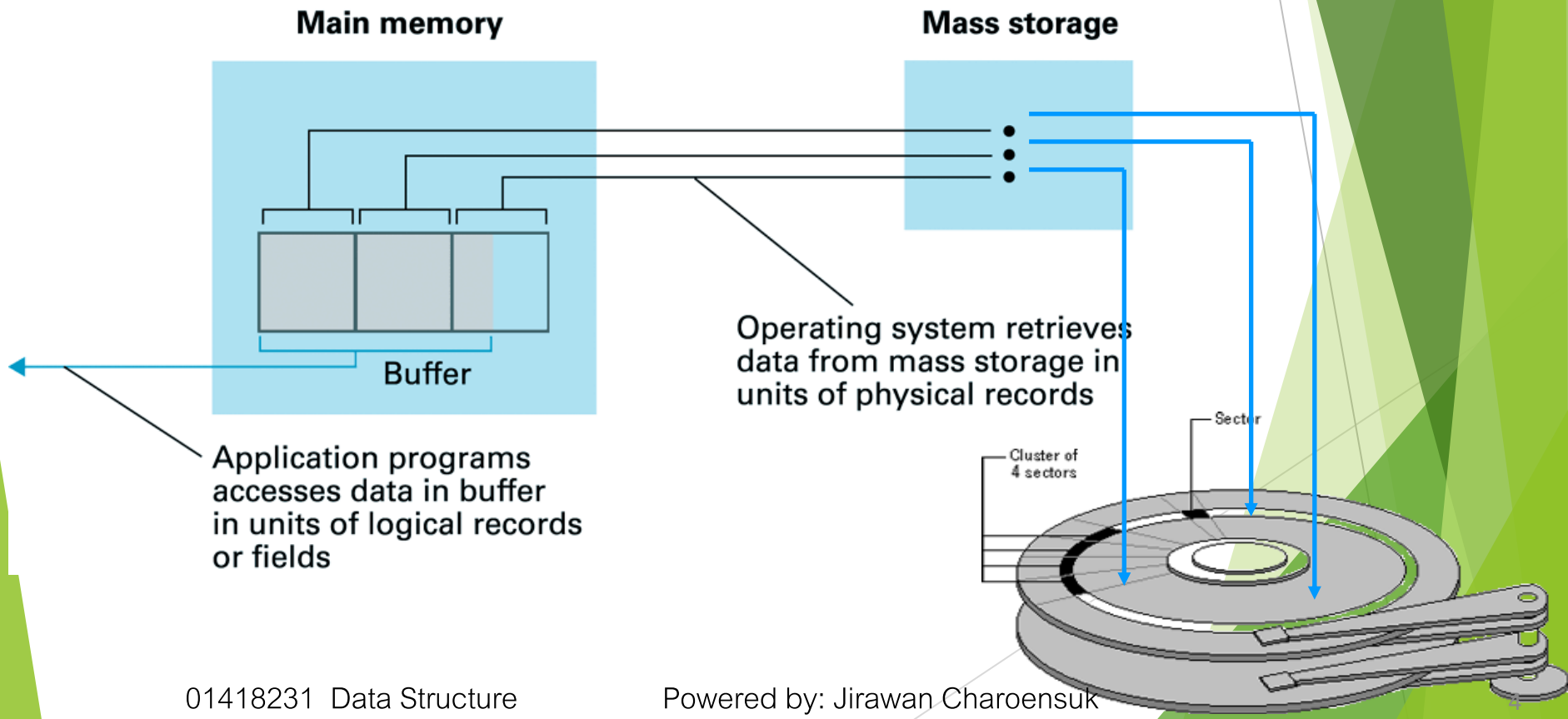
# Files, Directories & the Operating System

- **OS storage structure:**
  - conceptual hierarchy of directories and files



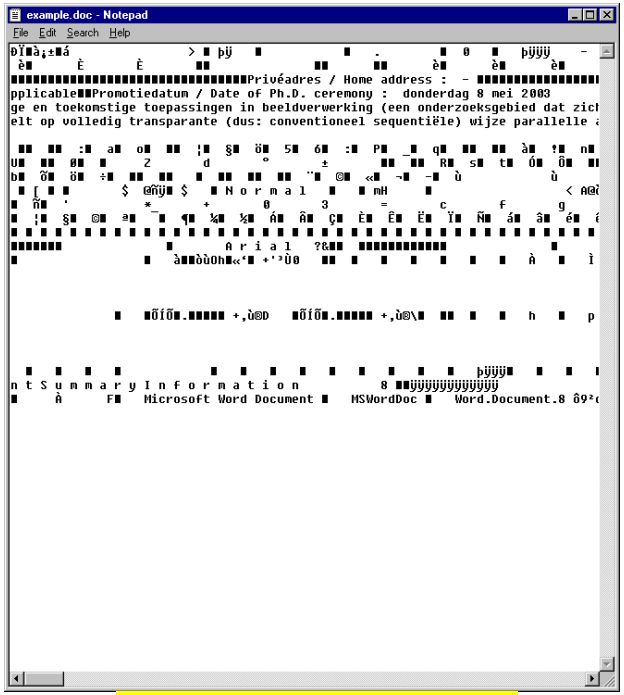
# Files: Conceptual vs. Actual View

- View at OS-level is conceptual
  - actual storage may differ significantly!

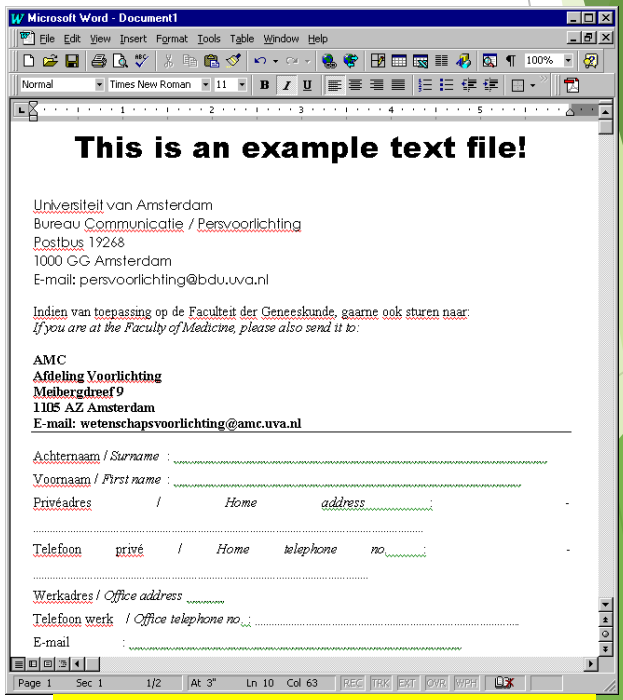


# Text Files

- Sequential file consisting of long string of encoded characters (e.g. ASCII-code)
  - But: character-string still interpreted by word processor!

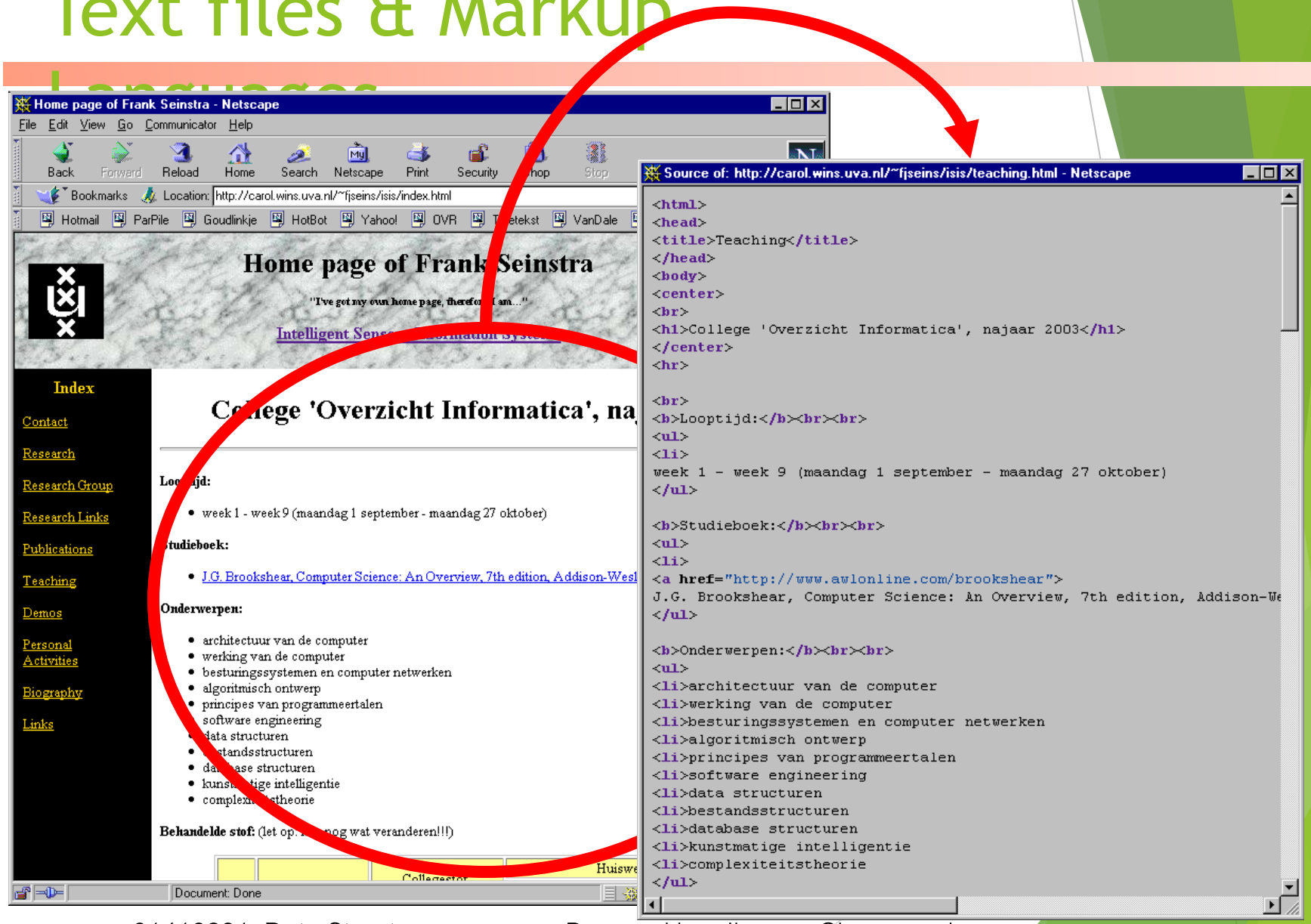



File in "Notepad"




Same file in "MS Word"

# Text files & Markup Languages



# Files

A file can be seen as

1. A stream of bytes (no structure), or
2. A collection of records with fields

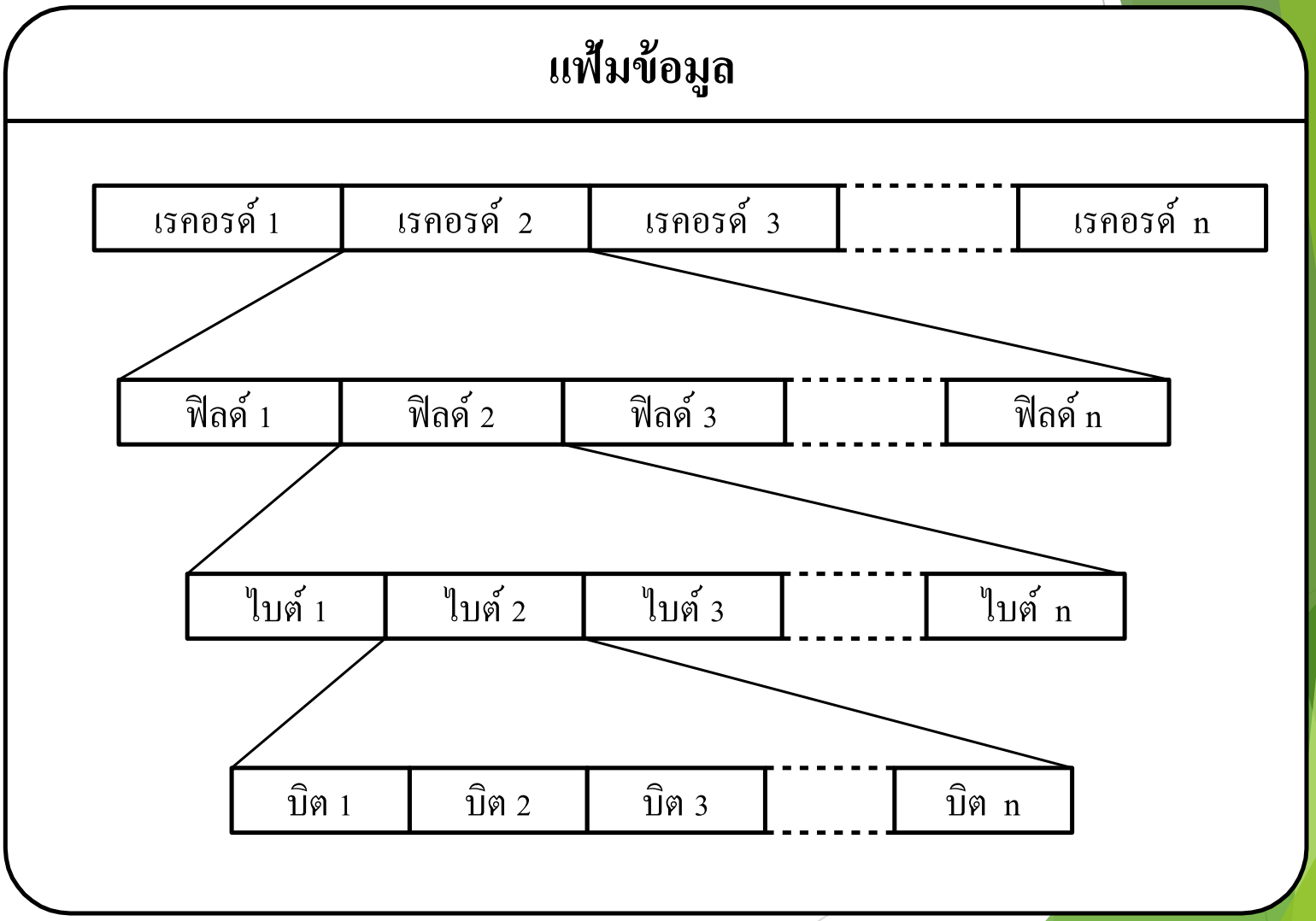
# A Stream File

- File is viewed as a sequence of bytes:

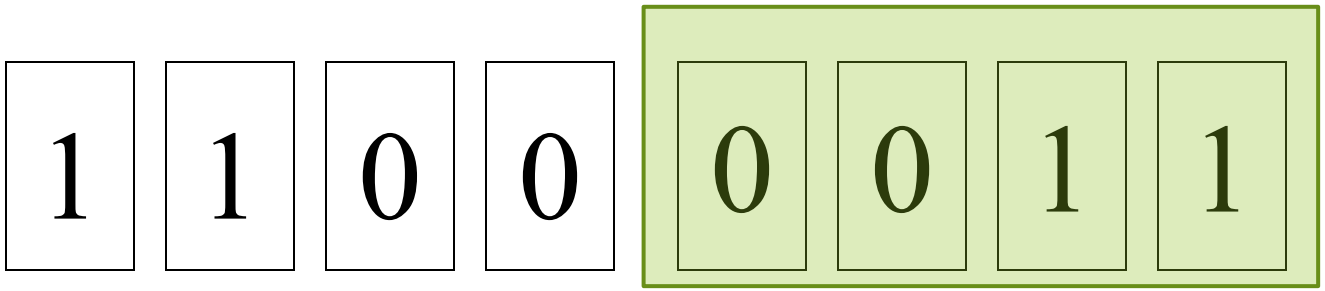
```
87359CarrollAlice in wonderland38180FolkFile Structures ...
```

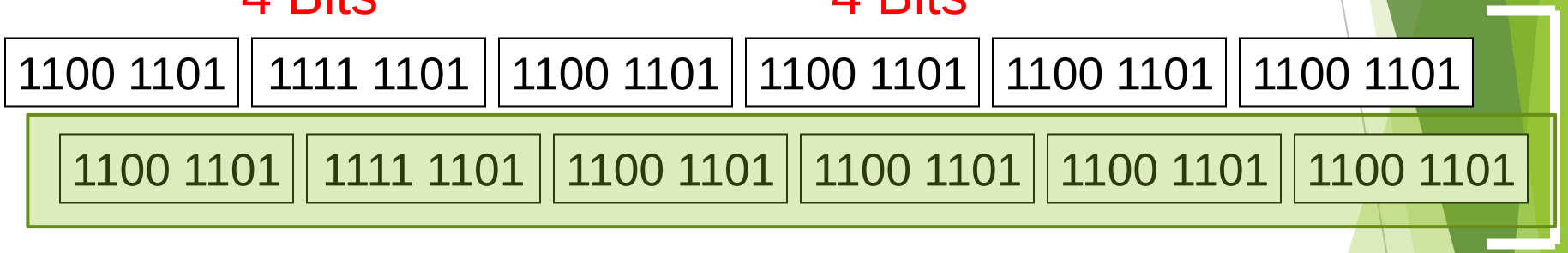
- Data semantics is lost:
  - there is no way to get it apart again.

# File Organization



# File Organization




= 1 Field

(Data Structure)

# Measurement

- KB (Kilo Byte) --> 1024 Bytes
- MB (Mega Byte) --> 1024 Kilo Bytes
- GB (Giga Byte) --> 1024 Mega Bytes
- TB (Tera Byte) --> 1024 Giga Bytes
- PB (Peda Byte) --> 1024 Tera Bytes

# Field and Record Organization

- Definitions
- Record: a collection of related fields.
- Field: the smallest logically meaningful unit of information in a file.
- Key: a subset of the fields in a record used to identify (uniquely) the record.

- e.g. In the example file of books:
  - Each line corresponds to a record.
  - Fields in each record: ISBN, Author, Title

# Example

| ISBN | book_name | author | publisher |
|---|---|---|---|
| 0001 | คู่กรรม | ทมยันตี | ดอกหญ้า |
| 0002 | บ้านทรายทอง | สมชาย | ดอกหญ้า |
| 0003 | ดาวพระศุกร์ | พลูโต | ดอกหญ้า |
| 0004 | พระเสาร์แทรก | สมหญิง | ดอกหญ้า |

# Record Keys

- Primary key: a key that uniquely identifies a record.
- Secondary key: other keys that may be used for search
  - Author name
  - Book title
  - Author name + book title
- Note that in general not every field is a key
  - (keys correspond to fields, or a combination of fields, that may be used in a search).

# Field Structures

- Fixed-length fields

```
87359Carroll    Alice in wonderland
38180Folk       File Structures
```

- Begin each field with a length indicator

```
058735907Carroll19Alice in wonderland
053818004Folk15File Structures
```

- Place a delimiter at the end of each field

```
87359|Carroll|Alice in wonderland|
38180|Folk|File Structures|
```

- Store field as `keyword = value`

```
ISBN=87359|AU=Carroll|TI=Alice in wonderland|
ISBN=38180|AU=Folk|TI=File Structures
```

# Field Structures

| Type | Advantages | Disadvantages |
|---|---|---|
| Fixed | Easy to read/write | Waste space with padding |
| Length-based | Easy to jump ahead to the end of the field | Long fields require more than 1 byte |
| Delimited | May waste less space than with length-based | Have to check every byte of field against the delimiter |
| Keyword | Fields are self describing. Allows for missing fields | Waste space with keywords |

# File Type

# Master file & Transaction file

# Master file

- Little frequency to update file
  - 1 year, 6 months
- Up to date -> using update file
  - The master file of bank's customer
    - ชื่อ ที่อยู่ หมายเลขบัญชี
  - The master file of student
  - The master file of bank's customer
  - The master file of book in library

# Transaction file

- More frequency to update file and continue to update
  - 1 hour,1 day, 1 week
- Uses to update Master file
  - The transaction file of employee (in-out of work)
  - The transaction file of reserve fright
  - The transaction file of sell-buy in Amazon



19

# Overview of File types




20

# Type of Data Processing

Batch Processing

Interactive Processing

# Batch Processing

- Collect transaction in one period before process data
  - 1 days, 1 month
- Used for save cost and time
  - Summary -> ยอดขายสินค้า
  - Report -> รายงานการใช้โทรศัพท์ของลูกค้า

# Interactive Processing

- Process suddenly , if transaction change
- Data want to update in real-time
- High cost
  - การฝาก-ถอนเงิน
  - การจองตั๋วเครื่องบิน

# File Operations

- Typical Operations:
  - Retrieve a record
  - Insert a record
  - Delete a record
  - Modify a field of a record

# Example

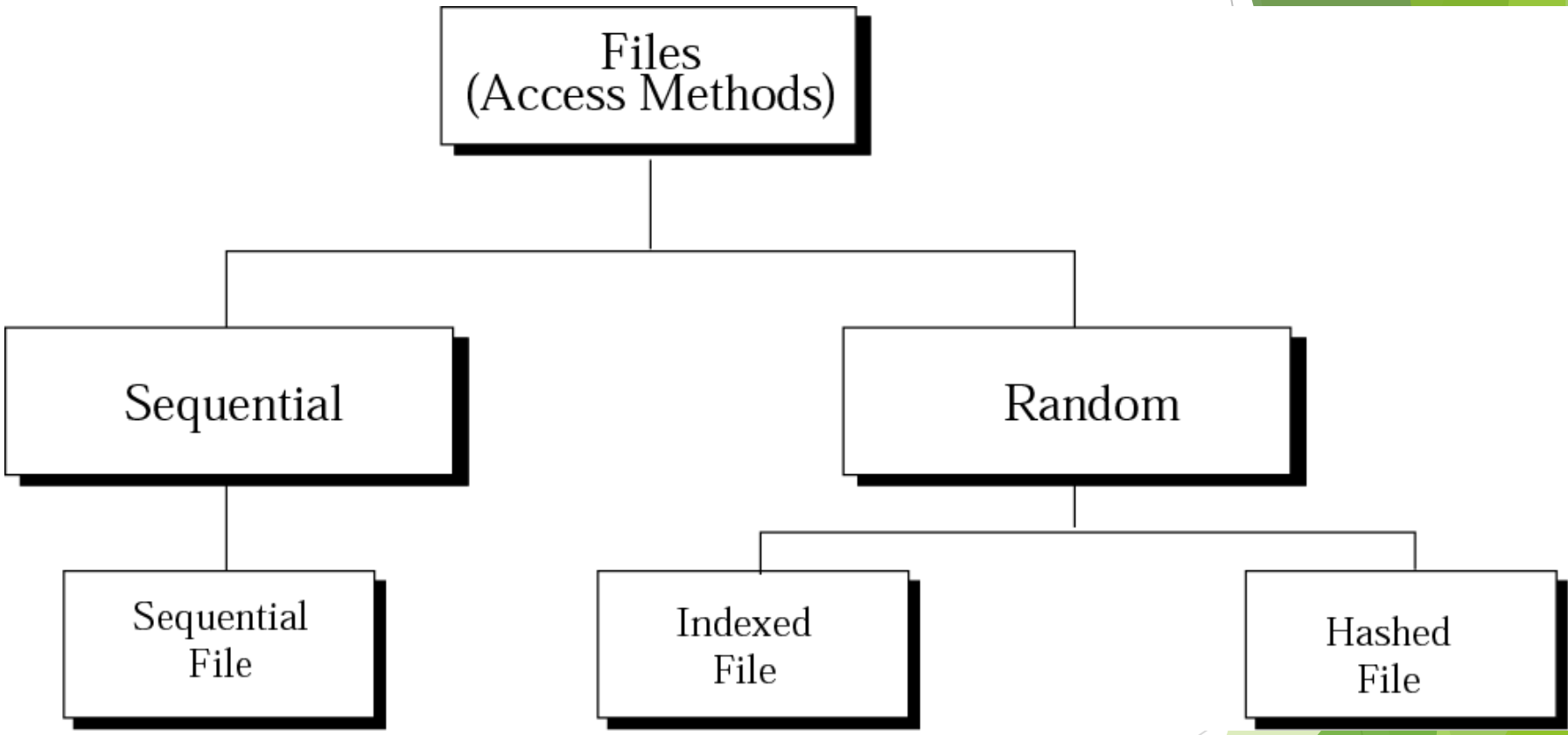
| ISBN | book_name | author | publisher |
|---|---|---|---|
| 0001 | คู่กรรม | ทมยันตี | ดอกหญ้า |
| 0002 | บ้านทรายทอง | สมชาย | ดอกหญ้า |
| 0003 | ดาวพระศุกร์ | พลูโต | ดอกหญ้า |
| 0004 | พระเสาร์แทรก | สมหญิง | ดอกหญ้า |

# Access Method

# Access methods

- A file is a collection of related data records treated as a unit.
- Files are stored in what are known as auxiliary or secondary storage devices.
- The two most common forms of secondary storage are optical and magnetic disks.
- A record in a file can be accessed sequentially or randomly.

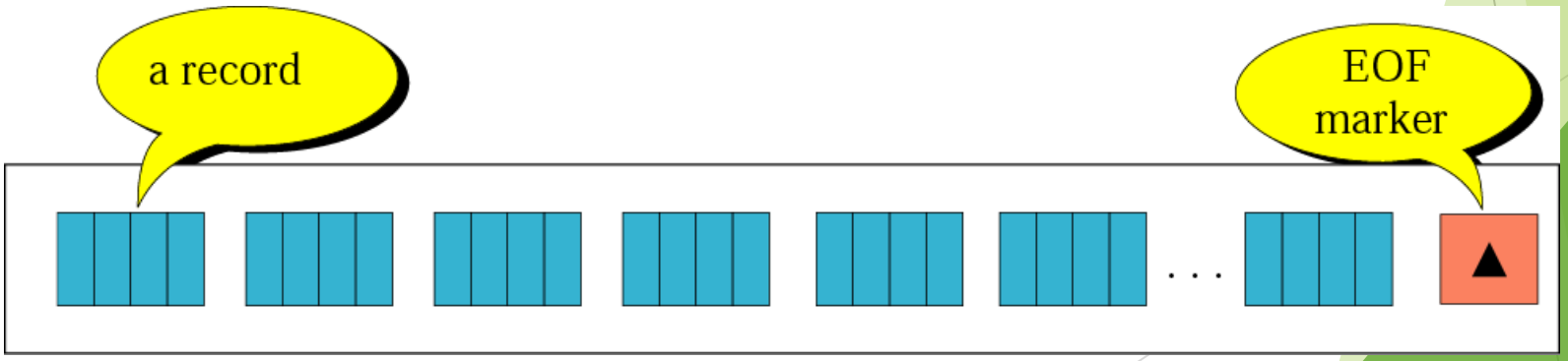
# Taxonomy of file structures



# Sequential file

# Sequential file

- Each record must be accessed sequentially, one after the other, from beginning to end.
- The update of a sequential file requires a new master file.
- An old master file, a transaction file, and an error report file.

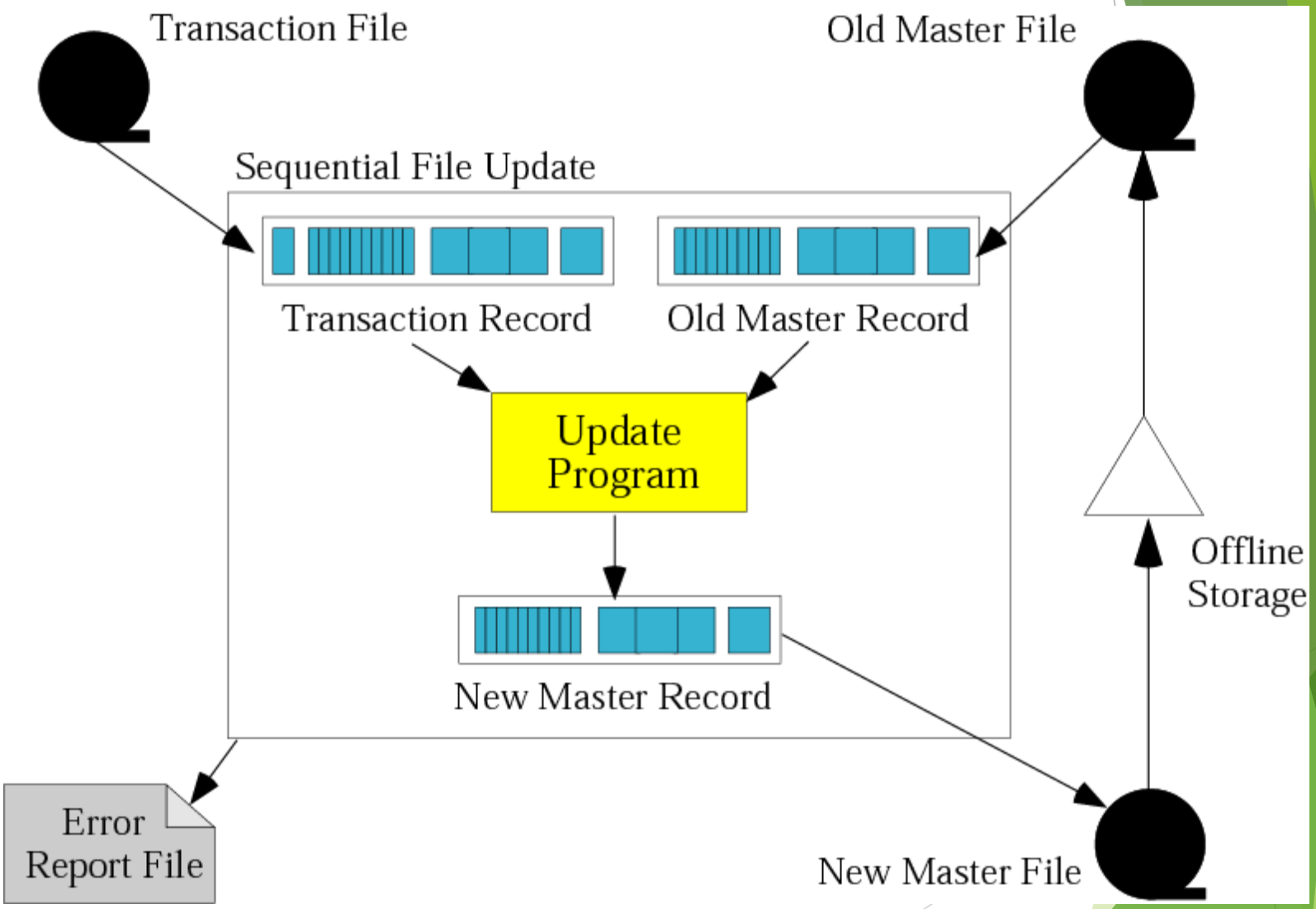


# Sequential file

**Program** **Processing records in a sequential file**

```
While Not EOF
{
    Read the next record
    Process the record
}
```

# Updating a sequential file



# Random file

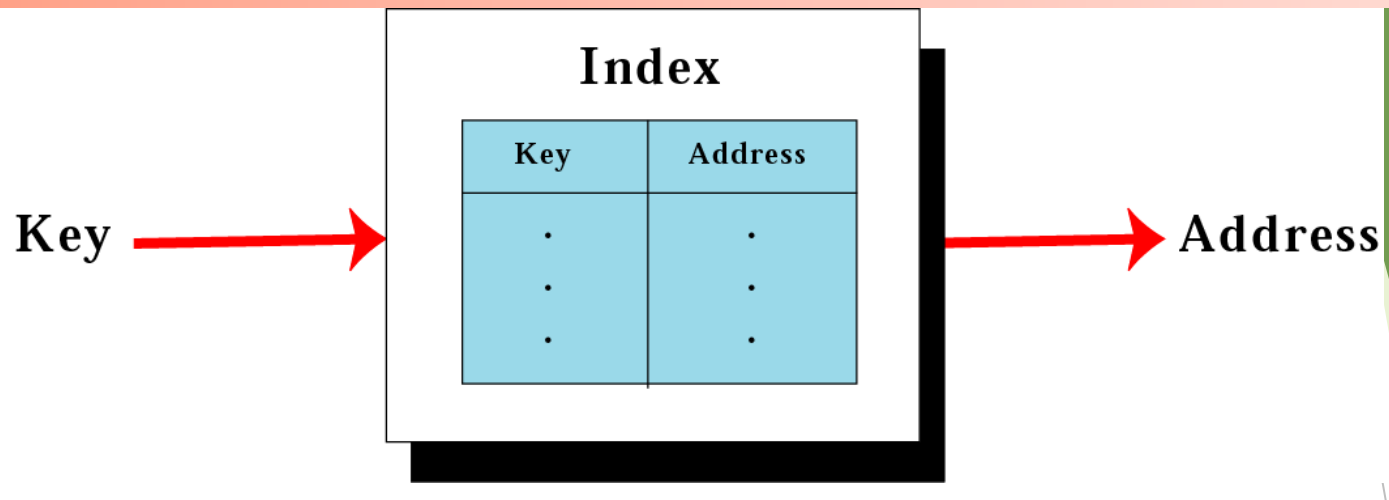
- Indexed files
- Hashed files

# Random access

- A record can be accessed without having to retrieve any records before it
- The address of the record must be known.
- For random access of a record,
  - an indexed file, consisting of a data file and an index, can be used.
- Using  index maps a key to an address, which is then used to retrieve the record from the data file.

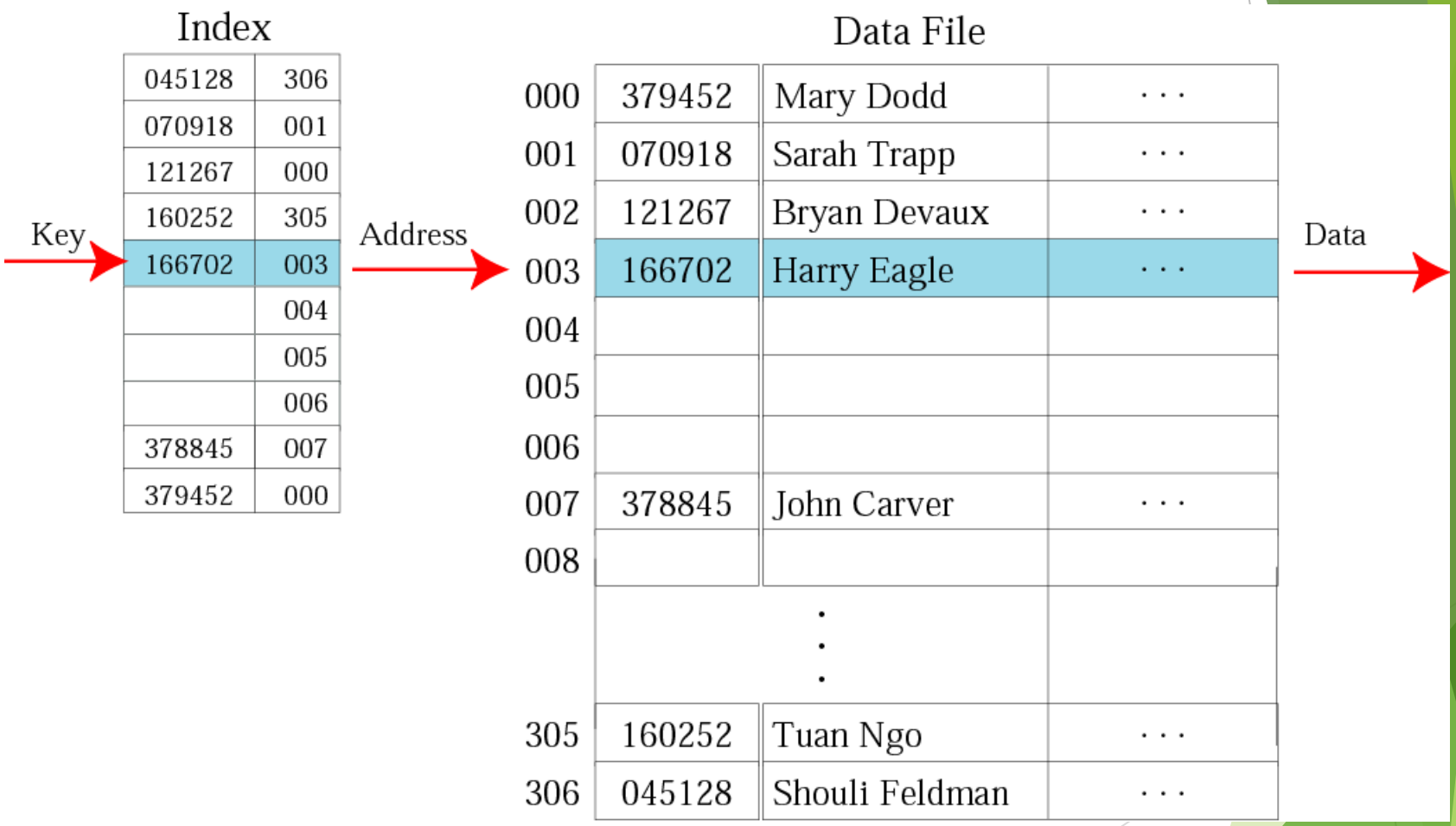
# Indexed files

# Mapping in an indexed file




- An indexed file is made of a data file,
  - which is a sequential file, and an index.
- The index itself is a very small file with only two fields:
  - the key of the sequential file and the address of the corresponding record on the disk.

# Logical view of an indexed file



# Indexed files

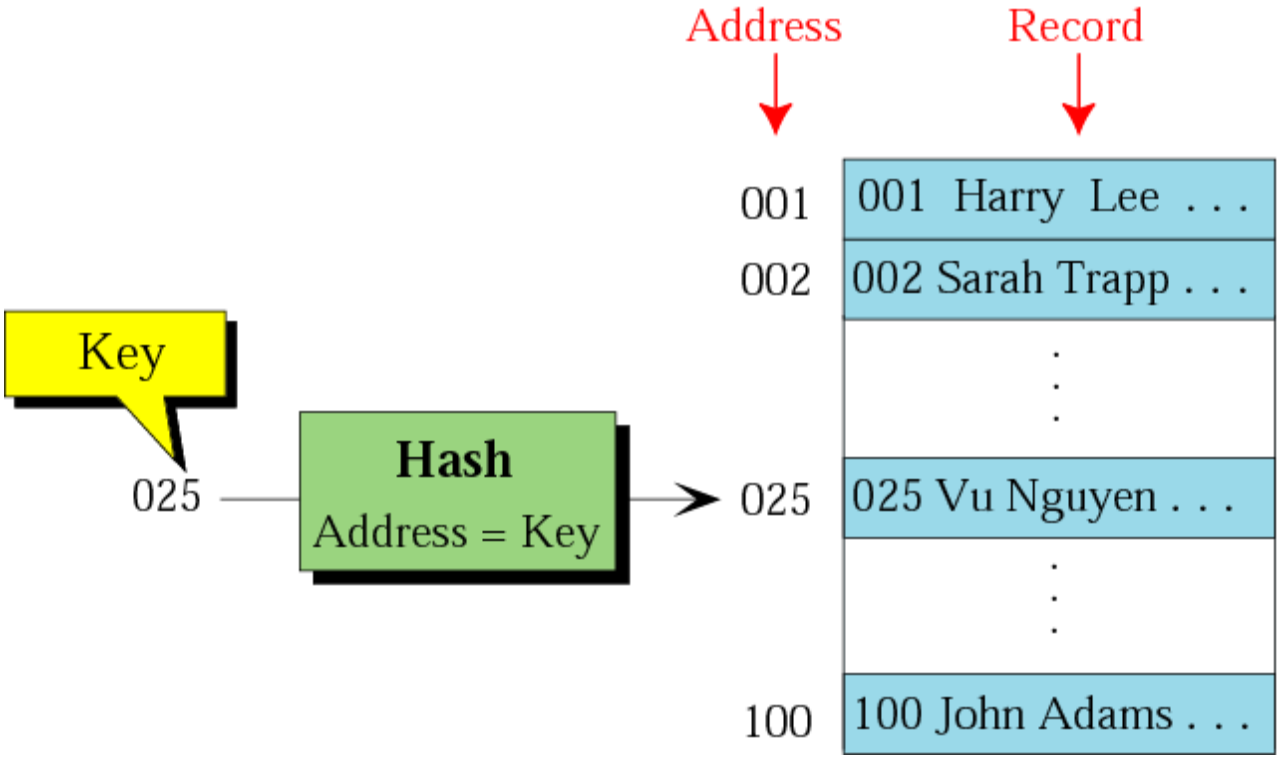
- An index file is made of a data file, which is a sequential file, and an index.
- Index – a small file with only two fields:
  - The key of the sequential file
  - The address of the corresponding record on the disk.
- To access a record in the file :
  1. Load the entire index file into main memory.
  2. Search the index file to find the desired key.
  3. Retrieve the address the record.
  4. Retrieve the data record. (using the address)

# Hashed files

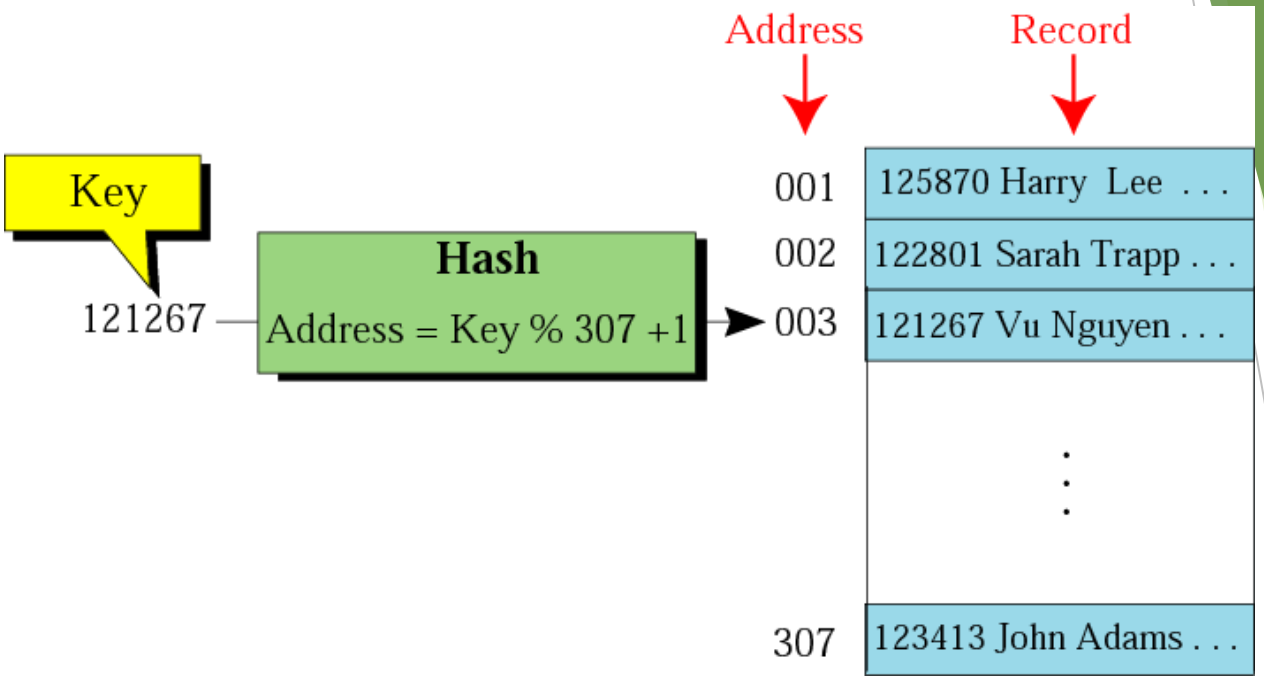
# Mapping in a hashed file

- A hashed file is a random-access file in which a function maps a key to an address.
- In direct hashing, the key is the address, and no algorithm manipulation is necessary.

# Direct hashing



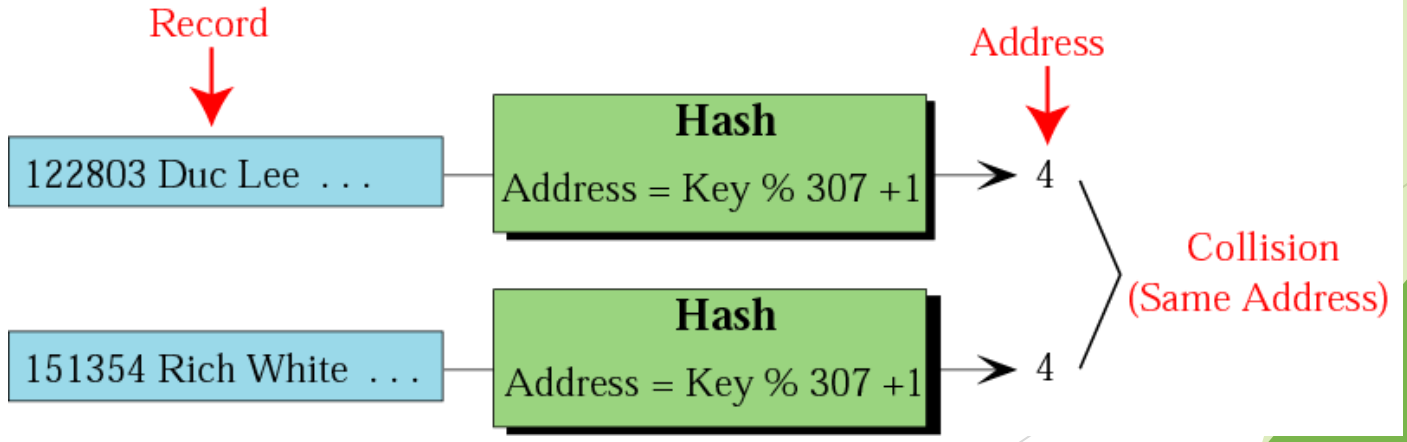
# Modulo division




- In **modulo division** hashing, the key is divided by the file size. The address is the remainder plus 1.

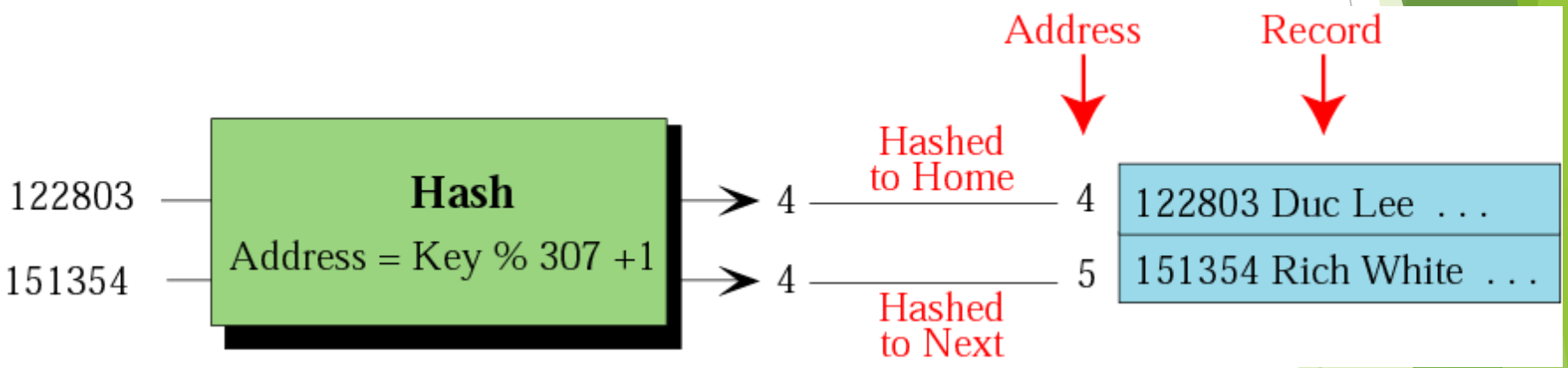
# Collision

- A collision is an event that occurs when a hashing algorithm produces an address for an insertion, and that address is already occupied.
- Collision resolution methods move the hashed data that cannot be inserted to a new address.

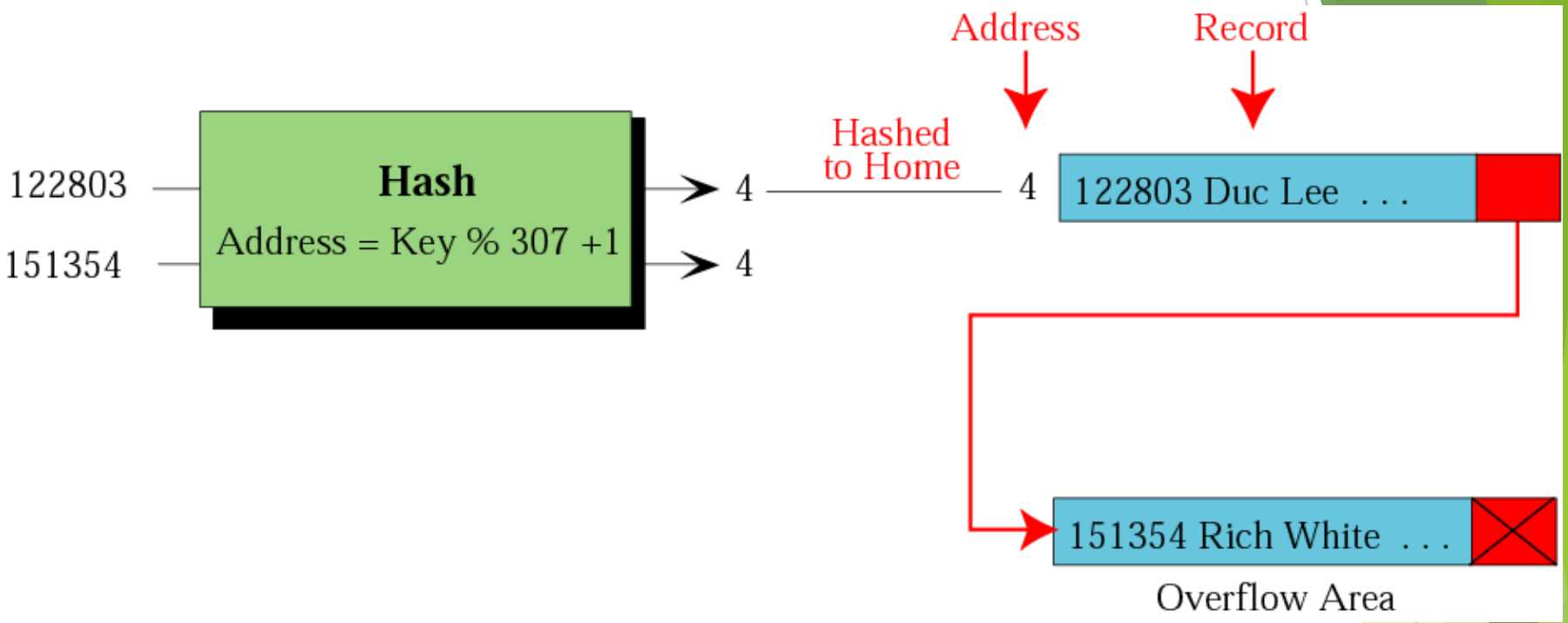


# Open addressing resolution

- The open addressing collision resolution method searches the prime area for an open address for the data to be inserted.

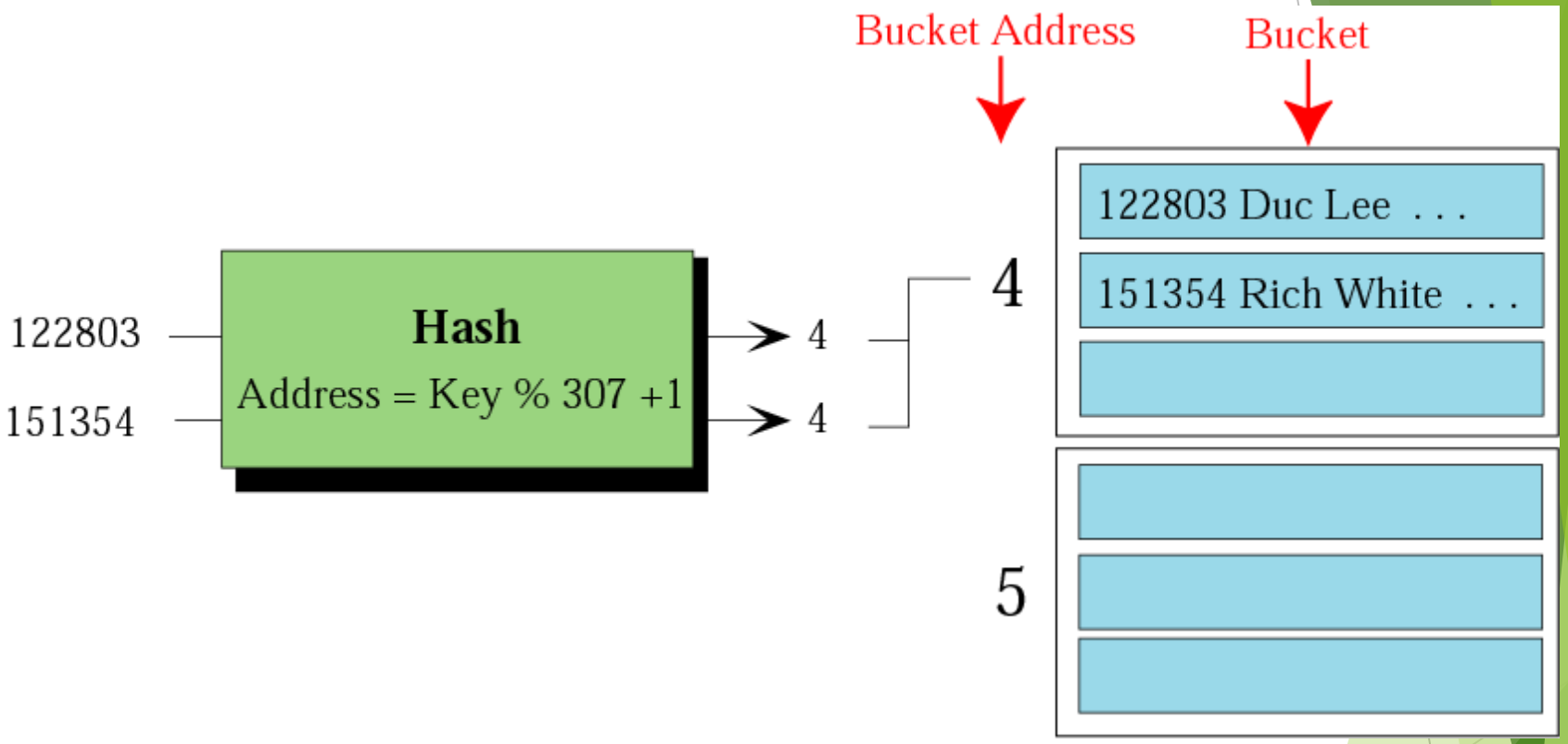


# Linked list resolution




- The linked list resolution method uses a separate area to store collisions and chains all synonyms together in a linked list.

# Bucket hashing resolution




- Bucket hashing is a collision resolution method that uses buckets, nodes that accommodate multiple data occurrences.

# Question

http://clipart-library.com/question-mark-gif.html