第50讲 文件的逻辑结构与访问方式

§5.2 File Organization and Access Methods

File Organization

- File organization refers to the logical structuring of the records as determined by the way in which they are accessed.

- The physical organization of the file on secondary storage depends on the blocking strategy and the file allocation strategy.

Criteria for File Organization

- **Rapid access**
  - Needed when accessing a single record.
  - Not needed for batch mode.
- **Ease of update**
  - File on CD-ROM will not be updated, so this is not a concern.

Criteria for File Organization

- **Economy of storage**
  - Should be minimum redundancy in the data.
  - Redundancy can be used to speed access such as an index.
- **Simple maintenance**
- **Reliability**

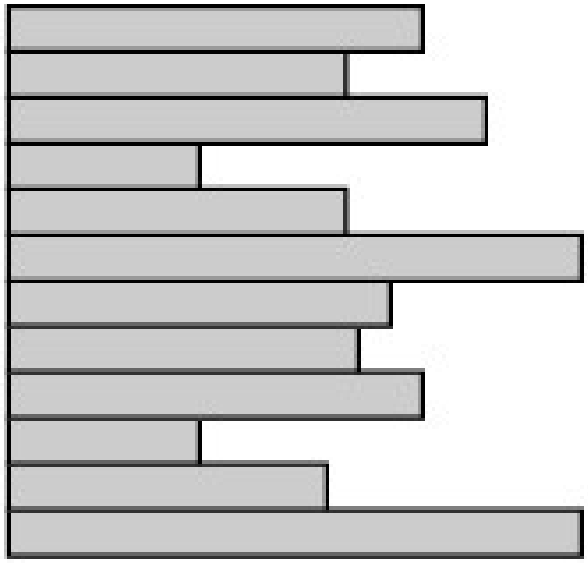
File Organization

- Pile file （堆文件）
- Sequential file （顺序文件）
- Indexed sequential file （索引顺序文件）
- Indexed file （索引文件）
- Direct, or Hashed file （直接或 hash 文件）

Pile(堆文件)

- Data are collected in the order they arrive.
- Purpose is to accumulate a mass of data and save it.
- Records may have different fields.
- No structure.
- Record access is by exhaustive search.

Pile



Variable-length records
Variable set of fields
Chronological order

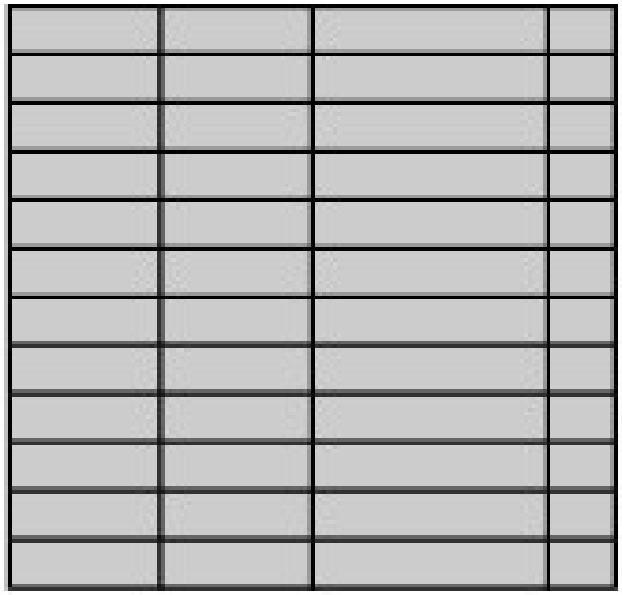
(a) Pile File

Figure 12.3 Common File Organizations

Sequential File

- Fixed format used for records.
- Records are the same length.
- All fields the same (order and length).
- Field names and lengths are attributes of the file.
- One field is the key filed
  - Uniquely identifies the record.
  - Records are stored in key sequence.

Sequential File



Fixed-length records
Fixed set of fields in fixed order
Sequential order based on key field

(b) Sequential File

Figure 12.3 Common File Organizations

Sequential File

- The only one that is easily stored on tapes as well as disk .
- Typically used in batch applications and the applications processing of all records.
- For querying and/or updating an individual record , it provides poor performance.
- New records are placed in a log file or transaction file.
- Batch update is performed to merge the log file with the master file.

Indexed Sequential File

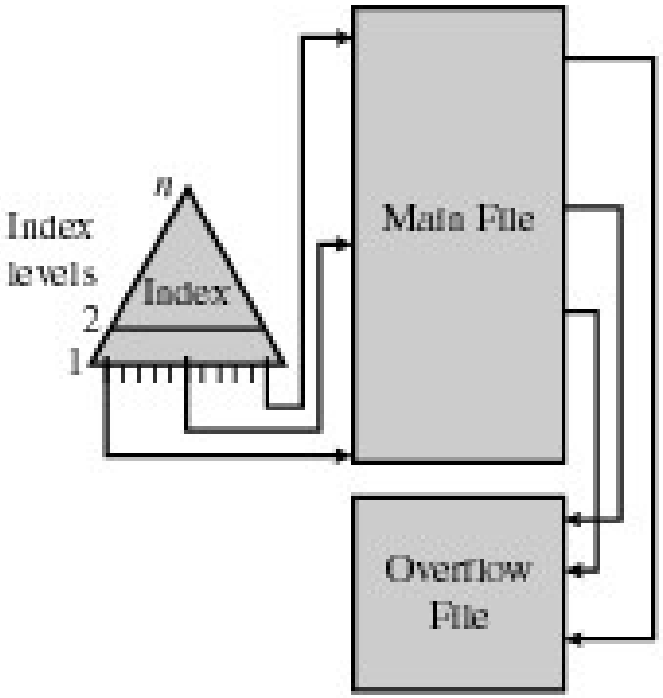
It maintains the key characteristic of the sequential file : records are organized in sequence based on a key field.

- **Add an index to the file to support random access , and an overflow file.**
- **Index file contains key field and a pointer to the main file , and provides a lookup capability to quickly reach the vicinity of the desired record.**
- **Overflow file is similar to the log file. It is integrated so that records in the overflow file are located by following a pointer from their processor record.**

Indexed Sequential File

- 索引顺序文件的每一条记录中都包含一个指向 overflow 文件的指针（该指针对应用程序透明）
  - New records are added to an overflow file
  - 如果新记录的前续记录在主文件（或 overflow 文件）中，则修改主文件（或 overflow 文件）中该前续记录指针，使之指向 overflow 文件中的新插入记录
  - The overflow is merged with the main file during a batch update.
  - Multiple indexes for the same key field can be set up to increase efficiency

Indexed Sequential File



(c) Indexed Sequential File

Figure 12.3 Common File Organizations

Sequential *vs.* Indexed Sequential

Example:

- A file contains 1 million records.
- On average 500,000 accesses are required to find a record in a sequential file.
- If an index contains 1000 entries, it will take on average 500 accesses to find the key, followed by 500 accesses in the main file. Now on average it is 1000 accesses.

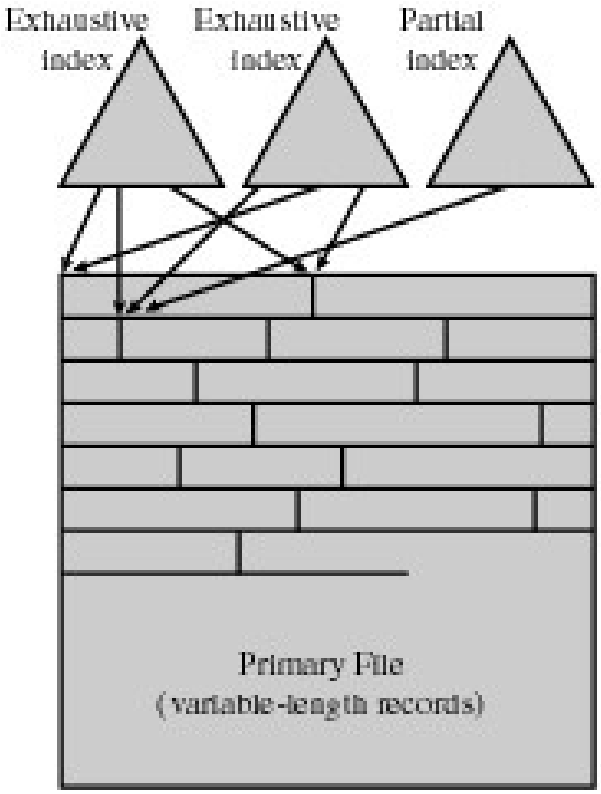
Sequential *vs.* Indexed Sequential

- 若一级索引文件包含 10,000 条记录，再为之建立二级索引，其中包含 100 条记录
- 在二级索引文件中平均检索 50 条记录，可以找到关键字
- 再在一级索引文件中平均检索 50 条记录找到指向主文件的指针
- 最后平均检索 50 条记录，找到目标记录，共计平均检索 150 条记录

Indexed File

- 顺序文件和索引顺序文件都只允许按记录的唯一关键字检索记录，这不符合某些应用需要按多个字段检索记录的要求
- 主文件的记录不必按关键字排序
- Uses multiple indexes for different key fields.
- May contain an exhaustive index(完备索引)that contains one entry for every record in the main file.
- May contain a partial index(部分索引)that contains entries to records where the field of interest exists.

Indexed File



(d) Indexed File

Figure 12.3 Common File Organizations

The Direct, or Hashed File

- Directly access a block at a known address.
- Key field required for each record.
- 利用 Hash 函数，根据记录的关键字计算记录的存储位置，并按关键字访问记录，提高记录的访问效率
- 常用于访问速率要求高、一次存取一条记录且记录为定长的文件。如文件目录、价格表、名单等文件记录的存取。