Soubory a databáze

Datový typ soubor

- **Soubor**
  - označuje množinu dat, která jsou kompletní k určitému zpracování a popisují vybrané vlastnosti reálných objektů
- **Záznam**
  - soubor se skládá ze záznamů, které popisují vybrané vlastnosti jediného objektu
- **Položka záznamu**
  - popisuje určitou vlastnost objektu
- **Klíč záznamu**
  - jedna z položek záznamu, která je pro každý záznam jedinečný (unikátní)
- **Návěští souboru**
  - sdružuje všechny informace potřebné k identifikaci souboru
    - jméno souboru
    - evidenční a pořadové číslo
    - datum vytvoření a doba platnosti
    - oprávnění přístupu
    - způsob práce se souborem
    - způsob kódování záznamů
    - délka souboru – počet záznamů
  - uloženo na začátku nebo konci souboru

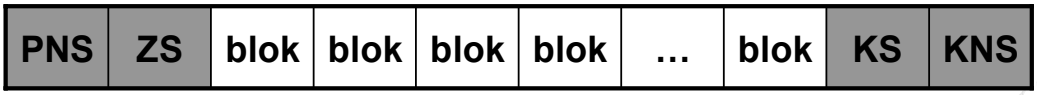
Operace se soubory

- **Rozdělení souborů podle způsobu přístupu k záznamům**
  - **sekvenční soubor**
    - přístupný je pouze aktuální záznam podle pořadí zpracovávání počínaje prvním záznamem
  - **soubor s náhodným přístupem**
    - záznamy jsou přístupné podle svého klíče
- **Operace nad souborem rozdělujeme podle úrovně pohledu**
  - **operace na úrovni záznamu**
    - operace s jednotlivými položkami záznamu
  - **operace na úrovni souboru**
    - operace se záznamy jako celky
  - **operace se soubory**
    - operace se soubory jako celky

http://kti.pef.czu.cz/Elvys

Implementace souborů

- **Sekvenční soubory**
  - reprezentovány posloupností **bloků** obsahujících **jeden nebo více datových záznamů**
  - této posloupnosti předchází blok s **návěštím souboru** (**PNS**) a **značkou souboru** (**ZS**)
  - soubor je ukončen blokem **konce souboru** (**KS**) a případně **koncovým návěštím** (**KNS**)

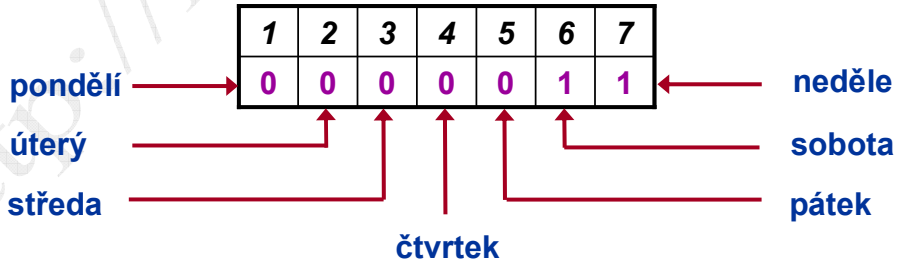


- **Soubory s náhodným přístupem**
  - záznamy v souboru jsou identifikovatelné pomocí **klíče**, který je součástí záznamu
  - **každý záznam** je adresovatelný – **zaujímá 1 blok**
  - **adresované soubory**
    - je definována **funkce** ***f(k)***, která jednoznačně přiřadí každému klíči *k* z množiny klíčů **K** ($k \in \mathbf{K}$) adresu ***adr* = *f(k)*** z intervalu adres **A** ($adr \in \mathbf{A}$) přidělených souboru na disku
  - **indexované soubory**
    - přiřazení adres daným klíčům je definováno **tabulkou indexů** – každému klíči je přiřazen adresní klíč, tj. adresa uložení záznamu

Kódování složek záznamu

- Na kódování složky záznamu má vliv nejen **typ hodnoty**, ale i **její rozsah**

  - **Pozičně kódované složky**
    - položka s předurčeným rozsahem *k* hodnot je reprezentována *k*-ticí bitů
    - hodnota v *i*-tém bitu reprezentuje *i*-tou hodnotu z *k* hodnot
    - **např**. dny v týdnu – reprezentace víkendu:



  - **Binární kombinační kódování**
    - položku s předurčeným rozsahem hodnot **1..*n*** zobrazíme *k*-ticí bitů, $k = \lceil \log_2 n \rceil$
    - každé hodnotě $h \in$ **1..*n*** přiřadíme jednu kombinaci 0 a 1 *k*-tice bitů
    - **např.** dny v týdnu:

001 – pondělí	101 - pátek
010 – úterý	110 - sobota
011 – středa	111 - neděle
100 – čtvrtek	

http://kit.pef.czu.cz/ENVS

- paměťově úspornější
- neumožňuje vytvářet kombinace hodnot

Položky s pevnou délkou
- položky s velkým předurčeným rozsahem hodnot se reprezentují základními datovými typy
- **integer**, **real**, **char**, **string**, **…**

Opakující se položky
- složkou záznamu nemusí být jen jediná hodnota, ale seznam hodnot, jehož délku předem neznáme
- řešením je omezení délky seznamu (např. poslední tři hodnoty)
- některé systémy ovládání souborů připouštějí záznamy proměnné délky – položka s proměnnou délkou je pak poslední položkou záznamu

Textové položky
- hodnoty textových položek jsou řetězce znaků libovolné délky
- obyčejně záznam obsahuje jedinou textovou položku nebo dokonce soubor obsahuje jediný záznam s textovou položkou
- textovou položku je obvykle nutné rozdělit do několika fyzických bloků

Odkazové položky
- hodnotou odkazové položky je odkaz na jiný záznam v souboru
- může mít formu klíče jiného záznamu, relativní pozice nebo fyzické adresy záznamu v souboru

Operace na úrovni souboru

Operace nad sekvenčním souborem

OPEN – otevření souboru
- soubor lze otevřít jako vstupní (pouze pro čtení) nebo jako výstupní (pouze pro zápis)

READ – přečtení aktuálního záznamu
- záznam je k dispozici až do čtení dalšího záznamu

WRITE – zápis záznamu do souboru

CLOSE – uzavření souboru
- pro správné uložení (i čtení) souboru je nutné soubor uzavřít

http://kti.pef.czu.cz/elvys

- **Operace nad souborem s náhodným přístupem**
  - **OPEN** – otevření souboru
    - otevření souboru pouze pro čtení, pouze pro zápis nebo pro čtení i zápis
  - **READ** – přečtení záznamu s daným klíčem
  - **WRITE** – zápis záznamu do souboru
  - **REWRITE** – zápis nové hodnoty záznamu s daným klíčem
  - **DELETE** – zrušení záznamu s daným klíčem
  - **CLOSE** – uzavření souboru

Operace se soubory

- Nad celými soubory zavádíme následující operace
  - **Konverze souboru**
    - kopírování souboru s případnými drobnými úpravami, souvisejícími s jinou vnitřní reprezentací souboru – **konverze pro** různá **vnější zařízení**, **komprimace** a **expanze**
  - **Řazení souboru**
    - seřazení záznamů souboru podle hodnoty klíče
    - řazení pomocí některé metod se sekvenčním přístupem – nejčastěji **Merge-sort**
  - **Slučování souborů**
    - sloučení dvou seřazených souborů v jediný soubor
  - **Aktualizace souboru**
    - promítání změn do všech nebo vybraných záznamů
    - generuje se nový soubor s aktualizovanými záznamy nebo se aktualizovaný záznam zapisuje na původní místo

Soubory a databáze

- Soubory dat jsou navrženy podle potřeb konkrétních programů, které je používají – s tím souvisí problémy:
  - **redundance a konzistence**
    - **redundance** = opakovaný (nadbytečný) výskyt některé informace ve více souborech
    - **konzistence** = opakovaně uložený údaj musí zachovávat stejnou hodnotu – popisuje vlastnost jediného objektu

http://ki.pef.czu.cz/Elvys

- **integrita**
  - všechna data uložená v souborech musí odpovídat vlastnostem popisovaných objektů reálného světa

- **obtížná dosažitelnost dat**
  - uživatelé potřebují získávat odpovědi na své dotazy ohledně stavu dat uložených v souborech
  - řeší se specializovanými programy – je nutné předem znát kostru dotazu

- **současný přístup více uživatelů**
  - systémy zpracování dat umožňují nebo potřebují současný přístup více uživatelů k souborům dat

- **ochrana dat proti zneužití**
  - není žádoucí, aby mohl kdokoliv provádět s daty jakékoliv operace nebo měl přístup ke všem uloženým informacím

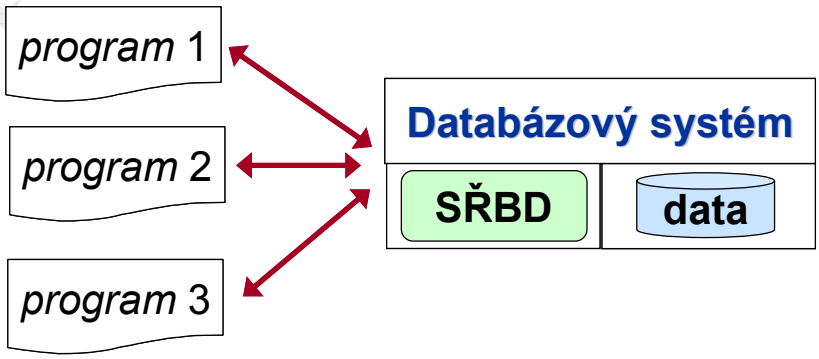
Databázové systémy

- Uvedené problémy se staly důvodem vzniku a rozvoje **databázových systémů** s následujícími vlastnostmi:
  - oddělené struktury aplikačních programů a datových souborů
  - přístup k datům je možný jen prostřednictvím programů databázového systému, ne přímo
  - dotazy nejsou pevné
  - je umožněn přístup více uživatelů současně
  - je vyřešena ochrana dat – přístupová práva

- **Databázový systém** se skládá z

  - **databáze (báze dat)** = množina vzájemně spolu souvisejících dat vztahujících se k určité problematice

  - **systému řízení báze dat (SŘBD)** = programový aparát pro popis a manipulaci s uloženými daty



http://kti.pef.czu.cz/Elvys

- **Databázový systém** zahrnuje:
  - **prostředky pro popis dat**
    - **DDL** (*Data Definition Language*)
  - **prostředky pro popis algoritmu práce s daty**
    - **DML** (*Data Manipulation Language*)
  - **prostředky pro popis ochrany přístupu k datům**
    - **DCL** (*Data Control Language*)
  - **dotazovací jazyk** pro uživatelské dotazy
    - **SQL** (*Structured Query Language*)
    - **QBE** (*Query By Example*)

http://kii.pef.czu.cz/Elvys