Classificação e Pesquisa de Dados

Aula 22

Operações sobre Arquivos; Arquivos Seqüenciais e Arquivos Seqüenciais Indexados

UFRGS INF01124

Revisão: conceitos

- Arquivo : coleção de registros lógicos
- Registro lógico : seqüência de campos ou atributos da entidade ou objeto sendo modelado
- Campo : corresponde a cada uma das informações que se deseja modelar a respeito da entidade ou objeto considerado

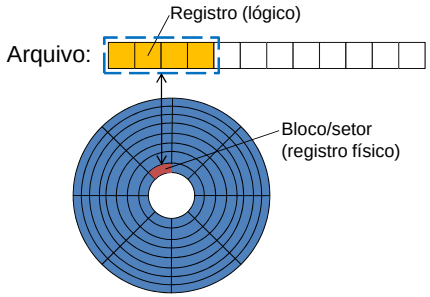
	ID	Nome	Salário
Registro 1:	00101	Leonardo	2.500,00
Registro 2:	00102	Aline	3.500,00
Registro 3:	00104	João	1.800,00
:	:	:	:
Registro *n*:	2193	Maria	4.500,01
	Campo 1	Campo 2	Campo 3

Revisão: conceitos

- Registro físico:
  - O armazenamento de um arquivo é feito, via de regra, por blocos de registros lógicos
  - Um bloco corresponde a quantidade de dados transferidos em um acesso simples
  - Um bloco de registros lógicos corresponde a um registro físico
  - Em cada leitura ou gravação é lido ou gravado um bloco e não apenas um registro lógico
  - Usualmente, o tamanho dos blocos é constante e coincidente com uma unidade de armazenamento do meio físico utilizado (ex.: Setores, trilhas em disco magnético)
  - Fator de bloco: número (inteiro) de registros lógicos por bloco

Revisão: conceitos

- Registro lógico versus registro físico:
  - Registro físico = setor/bloco (menor unidade de manipulação do dispositivo)
  - Estrutura de organização física do dispositivo normalmente difere da estrutura lógica!
  - Vários registros lógicos podem estar contidos em um bloco ou setor



Conseqüências:
- Vários registros lógicos são transferidos em uma única operação
- Organizar os arquivos de forma adequada, pode tornar o acesso às informações de um arquivo muito mais rápido
- Organização inadequada pode elevar o número de acessos

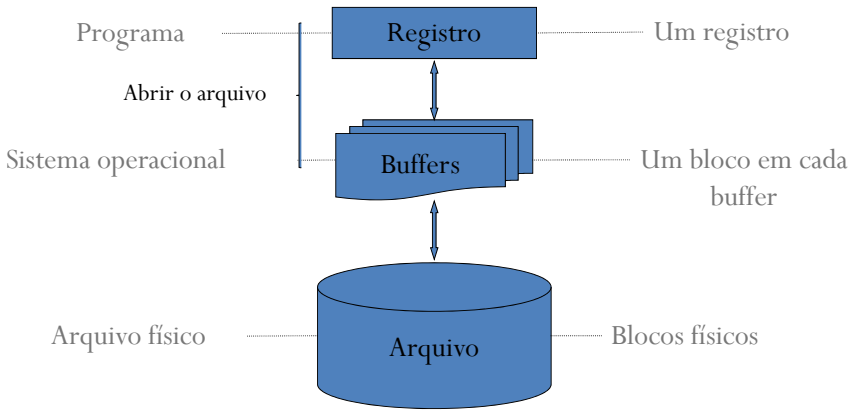
Organização de arquivos

- Registros lógicos não correspondem a um registro físico
- Aplicações precisam funcionar de modo independente dessa diferença
- É necessário realizar abstração entre esses dois tipos de registro
- **Organizar um arquivo significa relacionar (mapear) os requisitos de armazenamento de uma aplicação (dados/informações) com as características ou requisitos físicos de determinado dispositivo de armazenamento!**

Revisão: operações sobre arquivos

- Criação do arquivo
- Destruição do arquivo
- Abertura de um arquivo
- Inserção de registros
- Exclusão de registros
- Alteração de valores de campos
- Consulta a valores de campos (necessária para a execução das demais operações de alteração de registros)

Estrutura de acesso



Organização lógica de arquivos

- Não ordenada
  - Pesquisa serial
- Ordenada
  - Pesquisa seqüencial por chave de ordenação
  - Pesquisa binária por chave de ordenação
- Ordenada por freqüência de acesso
  - Pesquisa serial
- *Hashing*
  - Cálculo de endereço

Arquivos ordenados ou seqüenciais

- Uma classe importante de arquivos são os ordenados ou seqüenciais
- Um arquivo é seqüencial em relação a uma chave se as entradas no arquivo estão ordenadas por algum critério de ordenação em relação aquela chave
- Um arquivo seqüencial deve ter ao menos dois *buffers* para assegurar o paralelismo de leitura e de *fetch* do disco

Estruturas de representação

- Contigüidade física
  - Relacionamentos entre as entradas são representados pela disposição física ordenada dos registros por chave de busca
  - Posicional, implícita
- Encadeamento
  - Relacionamentos de ordenação entre os registros são representados por meio de apontadores simbólicos
  - Ligações explícitas

Organização de arquivos

- A forma de organização de arquivos (distribuição das entradas) deve levar em consideração as características da aplicação:
  - Volume de dados
  - Freqüência de atualizações
  - Freqüência de consultas
  - Natureza da chave (primária ou secundária)
  - Meio de armazenamento

Organização Física

- Contigüidade física
  - Mais natural
  - Acesso serial
  - Acesso direto se a chave for o número lógico do registro
- Lista encadeada
  - Acesso seqüencial
  - Inserção / remoção mais rápida do que no físico contíguo
- Índices
  - Acesso seqüencial, por pesquisa binária ou direto
  - Sempre por via indireta

Arquivo seqüencial - AS

- Caracterização
  - Registros fisicamente ordenados por uma chave primária ou por uma chave de ordenação

Número	Nome	Idade	Salário
1000	Ademar	25	900
1050	Afonso	27	500
1075	Carlos	22	1200
1100	Darci	25	1500
1300	Eber	39	500
1350	Genaro	19	650
1400	Helena	19	420
1440	Maria	21	900
1480	Ramon	22	1340
. . .	. . .	. . .	. . .

Arquivo seqüencial

- Indicação de uso
  - Memória de acesso seqüencial
  - Indicado para arquivos que sofrem manipulação por lotes (em *batch*)
  - Também para registros de tamanho variável
- Contra-indicação
  - Quando há mais do que uma chave
  - Quando exige-se respostas em tempo real
  - Aplicações com inserções/exclusões ao acaso (randômicas)

Acesso a um registro

Arquivo Seqüencial

- Serial ou seqüencial
- Aleatório

Acesso seqüencial a um registro

- Registros fisicamente armazenados de acordo com a seqüência na qual são solicitados (chave de acesso)
- Na maioria dos acessos o registro solicitado estará em memória (*buffer*) por pertencer ao mesmo bloco do seu antecessor
- Daí a importância da especificação de mais de um *buffer* para este tipo de arquivo

Acesso aleatório a um registro

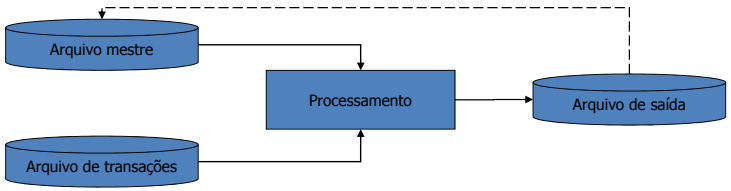
- Registro solicitado através de um argumento de pesquisa
- Casos possíveis:
  1. Argumento de pesquisa coincide com a chave de ordenação
     - Constatação rápida da ausência
     - Arquivo armazenado em dispositivo de acesso direto → possibilidade de busca via *pesquisa binária*
  2. Argumento de pesquisa não coincide com a chave de ordenação
     - Pesquisa serial (demorada)

Atualização

- Alteração de um registro
  - Situação normal:
    - Localiza → Lê → Altera campos → Grava
  - Situações especiais:
    - Alteração aumenta o tamanho do registro (registros de tamanho variável)
    - Alteração modifica o valor da chave de ordenação (corresponde a uma exclusão e uma posterior inclusão)

Atualização

- Inserções/exclusões de registros:
  - Arquivo mestre
  - Arquivo de transações
  - Execução de operação de intercalação
  - Arquivo de saída (novo mestre)



Arquivo seqüencial indexado

- Indicação de uso
  - No caso de arquivo seqüencial realiza-se o processamento completo do arquivo a cada ciclo de vida
  - Quando há necessidade de acessos aleatórios surge a necessidade de uma estrutura de acesso eficiente para a localização de um registro dado o argumento de pesquisa
  - Solução: criação de índice
- Uso prático: MySQL, dBase, Paradox, ou seja, bancos de dados (ver arquivos ISAM e suas extensões)

Caracterização

- Esta organização consiste essencialmente de um arquivo seqüencial acrescido de um índice, oferecendo acesso serial ordenado e acesso aleatório eficientes
- Além do arquivo seqüencial e do índice, esta organização ainda prevê uma área de extensão (ou área de *overflow*), utilizada para a implementação da operação de inserção de registros se for utilizado apenas o acesso aleatório

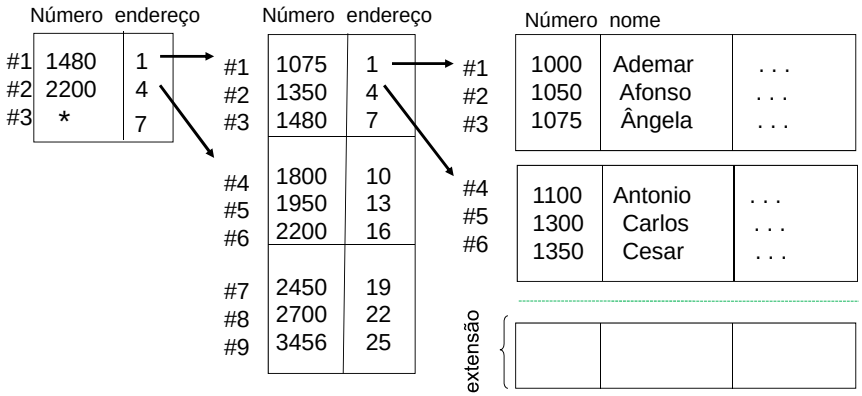
O que é um índice?

- Estrutura de acesso formada por uma coleção de pares do tipo <*chave, endereço*>, cada um deles associando o valor de uma chave de acesso a um endereço do arquivo na *memória secundária* (*disco*)
- Principais vantagens
  - Maior rapidez de busca
  - Permite múltiplos caminhos de acesso
  - Oferece maior facilidade de acesso a registros de tamanho variável

Tipos de Índices

- Índice primário
  - Índice associado a chave de ordenação
- Índices secundários
  - Não estão associados a chave de ordenação

Arquivo seqüencial indexado



Para tornar a pesquisa mais eficiente, o índice normalmente é estruturado em vários níveis (árvores B+, por exemplo)

Área de extensão

- Alternativas de implementação (mantendo o acesso seqüencial)
  - Associar um campo de elo em cada registro da área principal para conter o endereço da lista encadeada de seus sucessores alocados na área de extensão.
  - Em cada bloco de registros associar um campo de elo, destinado a conter o endereço da lista de extensões do bloco
    - Manutenção da seqüencialidade física dos registros dentro de cada bloco da área principal
    - Todos os registros da lista de extensão de um bloco possuem ordem maior do que todos os da área principal

Acesso a um registro

Arquivo Seqüencial Indexado

- Seqüencial: diretamente sobre a área de dados
  (para ordená-los são utilizados os algoritmos de classificação)
- Aleatório: é feito via índice
  (para acessar são utilizados os algoritmos de pesquisa)

Inserção de um registro

- Determinação do local onde deve ocorrer a inserção
  → busca no arquivo via índice.
  - Com área de extensão:
    - Inserção do registro no bloco selecionado ou na lista de extensão do seu sucessor na área principal.
  - Sem área de extensão:
    - As inserções são feitas em endereços, dentro de um mesmo bloco, liberados por exclusões ou reservados para este fim quando da geração do arquivo.
- Ambas as situações apresentam a necessidade de reorganização

Exclusão de um registro

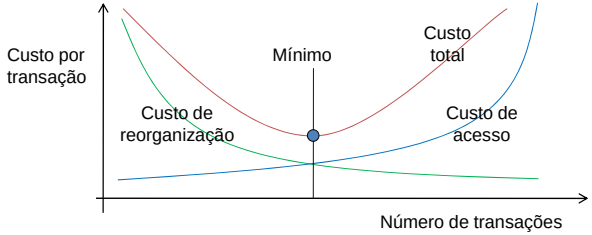
- Determinação do local onde deve ocorrer a exclusão
  → busca no arquivo via índice
- Colocação da marca de excluído
- Exclusão física com reordenação dos registros dentro do bloco ou reorganização da lista de *overflow*.

Alteração de um registro

- Determinação do local onde deve ocorrer a alteração → busca no arquivo via índice
- Procedimento normal
  - Quando a alteração não envolve a chave de ordenação e não aumenta o comprimento do registro
    (Lê → Altera → Grava)
- Procedimento diferenciado
  - Quando a alteração envolve a chave de ordenação ou altera o tamanho do registro
    (Lê → Exclui → Re-insere registro modificado).

Requisitos para reorganização

- Acessos freqüentes às áreas de extensão
- Necessidade de desconsideração dos registros excluídos
- Ponto de reorganização



Procedimento de reorganização

- Leitura exaustiva e transferência de todos os registros para uma nova área
- Todos os registros são colocados na área principal, ficando a área de extensão toda livre
- Registros excluídos desde a última reorganização são retirados fisicamente do arquivo
- Geração de um novo índice

Leitura recomendada

- Elmasri, Ramez; Navathe, Shamkant B. Record Storage and Primary File Organizations. In: **Fundamentals of database systems**. 2rd ed. Reading: Addison-Wesley, 1994. (ou suas traduções e versões mais recentes)
- Elmasri, Ramez; Navathe, Shamkant B. Index Structures for Files. In: **Fundamentals of database systems**. 2rd ed. Reading: Addison-Wesley, 1994. (ou suas traduções e versões mais recentes)