

Organização de arquivos

Projeto de arquivos
Prof. Allan Rodrigo Leite

Organização de arquivos

- Armazenamento de um pequeno volume de dados
 - Distribuição simples dos registros em um arquivo
 - Sem alguma estratégia para organização dos registros
 - Eficiente quando a frequência de acessos aleatórios não é elevada
 - Acesso aleatório é a recuperação de um registro em específico
- Armazenamento de um grande volume de dados ou aumento da complexidade dos acessos
 - Baixa eficiência no armazenamento e acessos aos registros
 - Requer técnicas sofisticadas para armazenamento e recuperação

Estratégias de organização de arquivos

- Diferentes cenários ou problemas com determinadas características requerem diferentes soluções para aumentar a eficiência
 - Arquivo sequencial simples
 - Arquivo sequencial ordenado
 - Arquivo sequencial-indexado
 - Arquivo indexado
 - Arquivo direto
 - Arquivo invertido

Definições sobre registros e arquivos

- Arquivo
 - Coleção de registros lógicos, representando um objeto ou entidade
- Registro lógico
 - Sequência de itens que representam campos ou atributos do registro
 - Atributo é uma propriedade constituída por nome, tipo e comprimento
 - Observação: o comprimento pode ser constante ou variável
- Registro físico
 - Armazenamento dos registros lógicos (leitura e gravação) em blocos
 - Tamanho do bloco coincide com uma unidade de armazenamento utilizada pelo meio físico (setores e trilhas de um hard-disk, por exemplo)
 - Cada bloco armazena um número inteiro de registros

Definições sobre registros e arquivos

- Chave
 - Sequência de um ou mais atributos de um registro
- Chave primária
 - Atributo que identifica exclusivamente cada registro do arquivo
- Chave secundária
 - Atributo utilizado para identificação (geralmente em índices)
 - Pode ter seu valor repetido em diferentes registros
- Chave de acesso
 - Chave utilizada para identificar registros em uma operação de leitura

Definições sobre registros e arquivos

- Argumento de pesquisa
 - Valor da chave de acesso em uma operação de leitura
- Chave de um registro
 - Valor de uma chave primária em um registro
- Chave de ordenação
 - Chave primária utilizada para estabelecer a sequência na qual devem ser dispostos (física ou logicamente) os registros de um arquivo

Arquivo sequencial simples

- Definição
 - Os registros são distribuídos em uma ordem arbitrária dentro do bloco
 - Isto é, os registros são dispostos um após o outro
 - Em geral a ordem pode ser a mesma da geração dos registros
- Vantagem
 - Simplicidade de implementação
- Desvantagem
 - Busca de registro por meio de acesso sequencial
 - Em casos de arquivos com muitos registros, esta busca é ineficiente

Arquivo sequencial ordenado

- Os registros estão dispostos ordenadamente
 - Sequência definida por uma chave primária (chave de ordenação)

Chave de ordenação

Localizar empregado com matrícula 1020

Chave de pesquisa: 1020

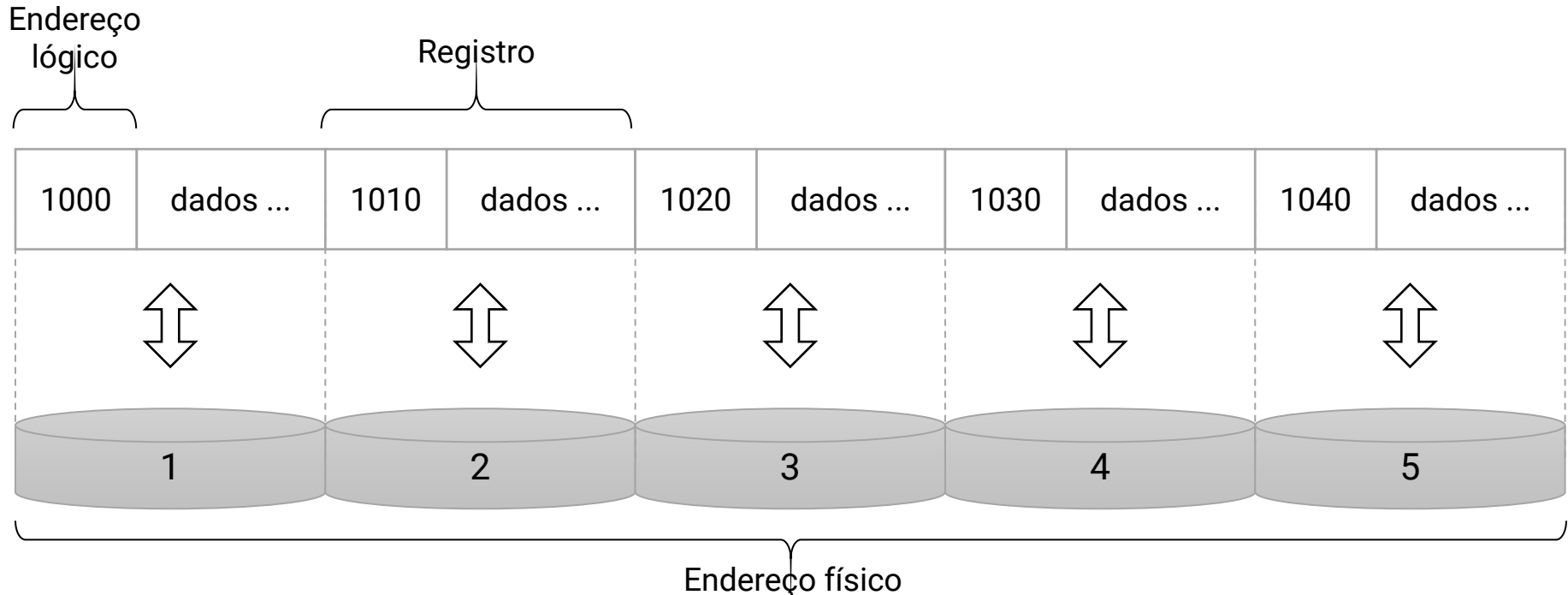
Matrícula	Nome	Data nascimento	Salário
1000	Ademar	11/02/1990	5000
1010	Clara	17/01/1985	7500
1020	Gerson	05/12/1988	6000
1030	Barbara	14/12/1992	4500

Arquivo: empregado

Atributos

Arquivo sequencial ordenado

- Estrutura do registro lógico e físico



Arquivo sequencial ordenado

- Principais características
 - Registros são gravados sequencialmente por suas respectivas chaves
 - Segue uma organização perfeita, tanto lógica quanto física
 - Registros possuem o mesmo formato
 - O valor de atributo está associado ao nome pela posição relativa no registro
 - A estrutura (layout) do registro é externa aos dados que ela descreve
 - A descrição é declarada nos programas por declarações de tipos e tamanhos
 - Campos alfanuméricos são dimensionados pelo tamanho máximo
 - Devido ao formato único para todas as ocorrências do registro
 - Portanto, pode ocorrer desperdício de posições de armazenamento

Arquivo sequencial ordenado

- Vantagens

- Operação de acesso sequencial eficiente quando:
 - Acesso a um registro cuja chave de acesso coincide com a de ordenação
 - Leitura dos registros do arquivo na sequência da chave de ordenação

- Desvantagens

- Operação de acesso (leitura) é ineficiente quando
 - Chave de acesso não coincide com a chave de ordenação
- Operação de gravação no arquivo pode requerer uma reorganização
 - Inserção, alteração e remoção de registros em um arquivo

Arquivo sequencial ordenado

- Operações de acesso e manipulação de registros
 - Operações de leitura
 - Acesso sequencial aos registros
 - Acesso aleatório a um registro
 - Leitura exaustiva (*full-scan*) dos registros
 - Operações de gravação
 - Inserção de um novo registro
 - Remoção de um registro existente
 - Alteração de um registro existente
 - Reorganização do arquivo

Operações em arquivo sequencial ordenado

- Acesso sequencial a um registro
 - Recuperação do registro que segue ao último acessado na sequência, segundo a chave de ordenação
 - Acesso eficiente quando
 - Registros fisicamente armazenados na sequência de acesso
 - Na maioria dos acessos, o registro desejado já estará na memória
 - Por pertencer ao mesmo bloco de seu antecessor
 - Exemplo: acessar os 3 primeiros empregados ordenados pela matrícula



Operações em arquivo sequencial ordenado

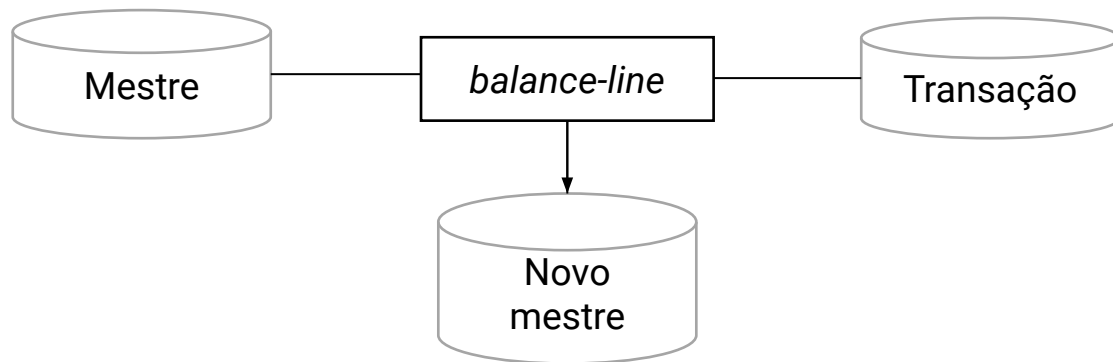
- Acesso aleatório a um registro
 - Baseado em um argumento de pesquisa
 - Não requer relação com a ordenação física do arquivo
- Cenários
 - Chave de pesquisa não coincide com chave de ordenação
 - Neste caso o acesso é sequencial
 - Chave de pesquisa coincide com chave de ordenação
 - Em mídia de acesso sequencial, a comprovação de registro não encontrado é mais rápida
 - Em mídia de acesso direto, usa-se pesquisa binária ou por interpolação para maior eficiência

Operações em arquivo sequencial ordenado

- Inserção de um novo registro
 - Utiliza uma técnica conhecida como *balance-line*
 - Inserir um único registro requer o deslocamento dos demais
- *Balance-line*
 - Gravações são realizadas em um arquivo temporário
 - Periodicamente estas operações são efetivadas no arquivo original
 - Para a efetivação é realizada a intercalação dos arquivos
 - Esta intercalação (temporário e principal) resultará em um novo arquivo

Operações em arquivo sequencial ordenado

- Procedimentos para inserção de um novo registro
 - Criar um arquivo de transação (temporário) com registros a gravar
 - Transação é uma sequência de operações para consistência dos dados
 - Conduz os dados de um estado consistente para outro estado consistente
 - Ordenar o arquivo temporária da mesma forma que o arquivo mestre
 - Intercalar os dois arquivos periodicamente
 - Gera-se um novo mestre com os registros reorganizados



Operações em arquivo sequencial ordenado

- Exclusão de um registro existente
 - Usa-se *balance-line* ou atributo adicional
 - Atributo adicional para indicar o estado do registro como excluído
 - Ou seja, com atributo adicional exclusão é apenas lógica
 - Acesso de leitura deve ignorar os registros marcados como excluídos
- Alteração de um registro existente
 - Usa-se *balance-line*
 - A alteração pode
 - Causar aumento do tamanho do registro
 - Modificar valor do campo usado como chave de ordenação

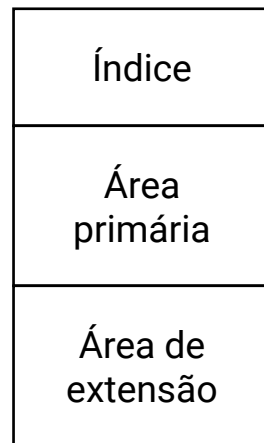
Operações em arquivo sequencial ordenado

- Leitura exaustiva dos registros
 - Manipula em paralelo os arquivos mestre e transação
- Reorganização do arquivo
 - Operação de intercalação entre os arquivos mestre e transação

Arquivo sequencial indexado

- Arquivo sequencial
 - Acesso aleatório
 - Sequência de acesso pode não coincidir com a ordenação física do arquivo
 - Torna-se um problema quando o volume de acessos aleatórios é grande
 - Neste caso, o ideal é o uso de uma estrutura de acesso mais eficiente
- Arquivo sequencial indexado
 - Arquivo sequencial com índice e área de extensão

Arquivo sequencial
indexado



Arquivo sequencial indexado

- Um arquivo sequencial indexado é constituído por 3 áreas
 - Área de índices
 - Arquivo sequencial criado pelo sistema
 - Cada registro deste arquivo estabelece uma divisão na área primária
 - Contém o endereço do início do segmento e a chave mais alta do mesmo
 - O sistema pode acessar de maneira direta um segmento da área de índices
 - Similar a busca por um capítulo de um livro a partir de seu índice
 - Área primária (principal)
 - Reservada para manter os registros de dados
 - Os registros são classificados em ordem ascendente pela chave primária
 - Área de excedentes (*overflow*)
 - Reservada para novos registros que não podem ser mantidos na área principal

Arquivo sequencial indexado

- Índice

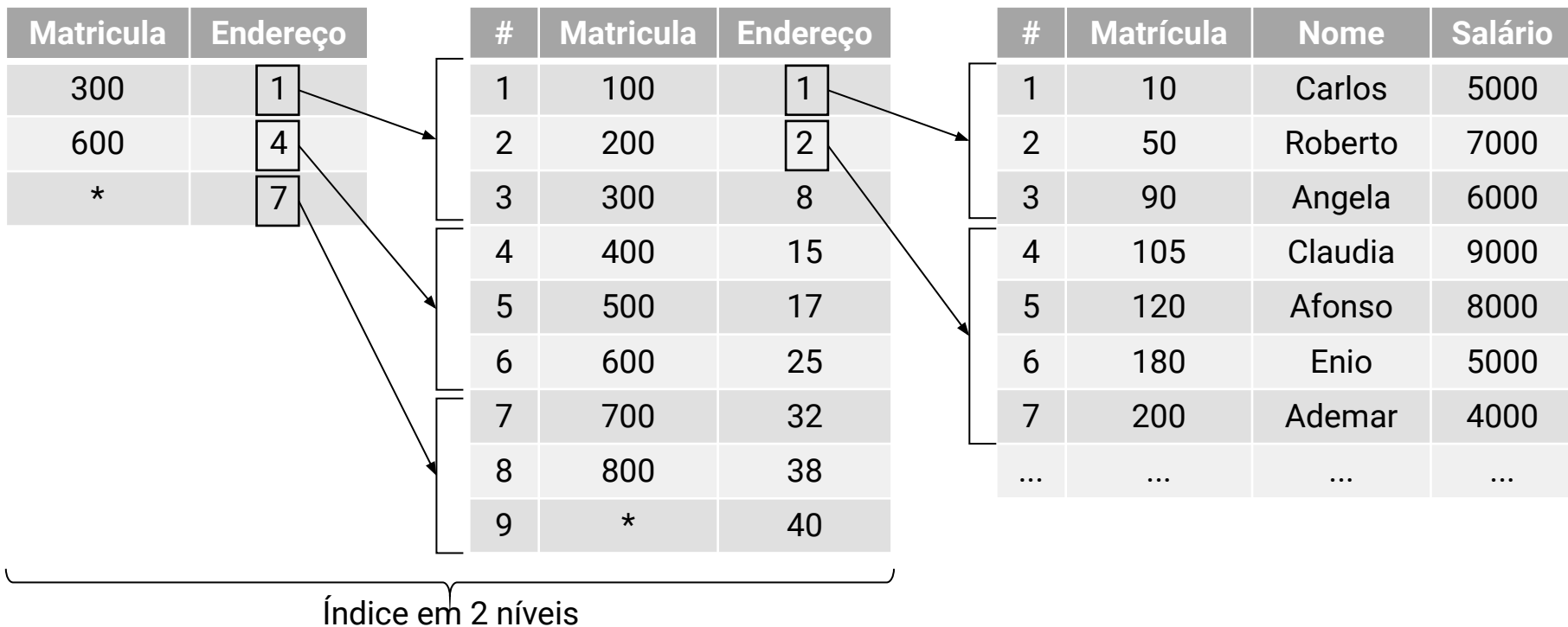
- Conjunto de entradas formado por uma coleção de pares chave-valor
- Cada entrada possui um valor de chave e um endereço do arquivo
- Deve ser especificado um índice para cada chave de acesso
- Permite uma rápida localização do endereço de um registro
 - A partir de um argumento de pesquisa

Cada entrada identifica
um bloco do arquivo {

Início do bloco	
Valor da chave	Endereço
100	1
200	4
300	8
400	13
500	18

Arquivo sequencial indexado

- O índice pode ser organizado em múltiplos níveis



Arquivo sequencial indexado

- Área de extensão
 - Contém os registros inseridos após a criação do arquivo principal
 - Extensão da área principal de dados do arquivo
 - Não é viável a realizar inserção de registros como em arquivo sequencial
 - Os registros podem mudar de endereço
 - Isto exige alterações das entradas dos índices

Arquivo sequencial indexado

- A área de extensão pode ser implementado de dois modos
 - Modo 1: cada registro da área de extensão possui um encadeamento indicando o seu antecessor ou sucessor
 - Modo 2: usar um atributo para encadeamento de cada bloco de registro contendo a lista de extensões do bloco
- Podem existir várias áreas de extensão em um mesmo arquivo
 - Uma para cada bloco ou grupo de blocos adjacentes
 - Uma ou mais áreas adicionais usadas sempre que ocorre uma inserção em um bloco cuja respectiva área de extensão já está cheia

Operações em arquivo sequencial indexado

- Operações de leitura
 - Acesso sequencial
 - Direto sobre a área de dados e extensão sem usar o índice
 - Acesso aleatório
 - Uso do índice para obter o endereço do próprio registro ou do seu bloco
 - Neste último caso, é necessária uma busca dentro do bloco
 - Também deve incluir mais acessos referentes à área de extensão
 - Leitura exaustiva (*full-scan*)
 - Igual ao acesso sequencial

Operações em arquivo sequencial indexado

- Operações de gravação
 - Inclusão
 - Usa as áreas de extensão
 - Exclusão
 - Pode ser utilizada a técnica de atributo adicional (exclusão lógica)
 - Alteração
 - Pesquisa-se o registro no arquivo
 - Se a alteração não afetar a chave de ordenação, o registro é sobrescrito
 - Do contrário, usa-se as operações de exclusão e inclusão

Operações em arquivo sequencial indexado

- Reorganização
 - O desempenho das operações (leitura e gravação) é degradado à medida que ocorre novas inclusões e exclusões de registros
 - A reorganização do índice deve ser realizada periodicamente para
 - Excluir (física e lógica) os registros excluídos
 - Sanear da área de extensão
 - Após a reorganização, um novo índice deve ser gerado
 - O intervalo de tempo entre cada reorganização deve ser estabelecido
 - A reorganização não deve ser realizada quando o arquivo estiver em uso
 - Exemplo: ao atingir mais que 75% de uso da área de extensão

Arquivo sequencial indexado

- Principais características
 - Permite acesso aleatório satisfatório
 - Permite acesso sequencial eficiente pela chave primária
 - Exemplo: impressão de relatório de todo estoque de um armazém
 - Facilita a inserção e exclusão de registros pelo uso da área de extensão

Arquivo indexado

- Motivação

- Para oferecer um acesso sequencial eficiente, os arquivos sequenciais ordenados requerem que os registros fisicamente ordenados
- Isto dificulta a inserção de um registro e exigindo
 - Utilização de áreas de extensão
 - Efetivação de reorganizações periódicas
- A manutenção da sequência dos registros torna-se inviável quando
 - Frequência de acessos sequenciais for baixa
 - Frequência de acessos aleatórios for alta

- Definição

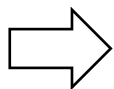
- Existência de um ou mais índices para acesso aos registros
 - Não há qualquer compromisso com a ordem física dos registros
 - Considera a possibilidade de acesso por qualquer atributo do registro

Arquivo indexado

- Suporte a múltiplos índices
 - Podem existir um índice para cada chave de acesso aos registros
 - O índice tem um conjunto de entradas ordenadas pelas chaves de acesso
 - Permite uma busca mais eficiente e o acesso sequencial ao arquivo
 - Cada entrada do índice contém o valor do atributo e um ponteiro ao endereço físico do registro
 - Não há entradas cujos ponteiros direcionam para blocos de registros
- Classificações de índices
 - Exaustivo: quando possui uma entrada para cada registro do arquivo
 - Seletivo: uma entrada para cada subconjunto de registros

Índice exaustivo

Entrada	Matrícula	Endereço
1	1000	301
2	1010	302
3	1020	303
4	1030	304
5	1040	305
6	1050	306



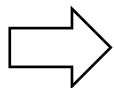
Endereço	Matrícula	Nome	Data nasc	Depto	Salário
301	1000	Ademar	11/02/90	A	5000
302	1010	Roberto	17/01/85	B	7500
303	1020	Gerson	05/12/88	A	6000
304	1030	Ieda	18/05/63	C	9000
305	1040	Bernardo	14/12/92	C	4500
306	1050	Angela	15/02/95	C	6500

Índice exaustivo (primário)

Área de dados

Índice seletivo

Entrada	Matrícula	Endereço
1	1000	301
2	1010	302
3	1020	303
4	1030	304
5	1040	305
6	1050	306



Endereço	Matrícula	Nome	Data nasc	Depto	Salário
301	1000	Ademar	11/02/90	A	5000
302	1010	Roberto	17/01/85	B	7500
303	1020	Gerson	05/12/88	A	6000
304	1030	Ieda	18/05/63	C	9000
305	1040	Bernardo	14/12/92	C	4500
306	1050	Angela	15/02/95	C	6500

Índice exaustivo (primário)

Área de dados

Depto	Entrada
A	1, 3
B	2
C	4, 5, 6

Índice seletivo (departamento)

Salário	Entrada
5000	1, 5
6000	3, 4
7500	2, 6

Índice seletivo (salário)

Operações em arquivos indexados

- Operações de leitura
 - Acesso sequencial
 - Utiliza-se o índice apropriado cuja identificação é simplificada, pois as entradas dos índices são ordenadas
 - Neste caso, a memória mantém um bloco do índice, reduzindo o número de leituras ao disco (memória secundária)
 - Acesso aleatório
 - Requer uma busca sobre os índices que suportam a chave de acesso
 - Leitura exaustiva (full-scan)
 - São realizados sucessivos acessos sequenciais sobre o índice exaustivo

Operações em arquivos indexados

- Operações de gravação
 - Inclusão
 - O registro pode ser armazenado em qualquer endereço disponível
 - Os seus pares são inseridos nos índices correspondentes
 - Para o tratamento dos índices é utilizada uma estrutura chamada Árvore B
 - Exclusão
 - Liberada a área de dados ocupada pelo registro
 - Removidas as entradas nos índices correspondentes ao registro
 - Alteração
 - Primeiro busca-se o registro pela chave de acesso
 - Em seguida os atributos são alterados e gravados na mesma posição

Arquivo indexado

- Vantagens
 - Operação de inserção mais eficiente
 - Possibilidade de acessos aleatórios a partir dos índices
- Desvantagens
 - Acesso sequencial ineficiente
 - Necessidade de manutenção de um ou mais índices
 - Inserções ou alterações envolvendo atributos associados aos índices

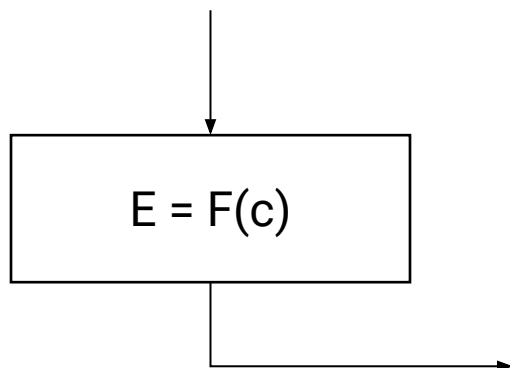
Arquivo direto

- Motivação
 - Acesso rápido aos registros especificados por argumentos de pesquisa, sem percorrer uma estrutura auxiliar (índice)
- Definição
 - Ao invés de um índice, é utilizada uma função (*hashing*) que calcula o endereço do registro a partir do valor da chave do registro

Arquivo direto

Argumento de pesquisa

c = 1040



Endereço	Matrícula	Nome	Data nasc	Depto	Salário
301	1000	Ademar	11/02/1990	A	5000
302	1010	Roberto	17/01/1985	B	7500
303	1020	Gerson	05/12/1988	A	6000
304	1030	Ieda	18/05/1963	C	6000
305	1040	Bernardo	14/12/1992	C	5000
306	1050	Angela	15/02/1995	C	7500

Onde:

E = Endereço

F = Função matemática (*hashing*)

c = Chave primária

Arquivo direto

- Abordagem similar ao arquivo indexado
 - Em ambos os casos o acesso aleatório é eficiente
- Diferenças para o arquivo indexado
 - No arquivo indexado o endereço é independente do valor da chave
 - No arquivo direto não são previstos acessos seriais

Funções para cálculo do endereço

- Funções determinísticas
 - Dada qualquer chave de acesso, sempre gera um único endereço
 - Em termos práticos não despertam maiores interesses
- Funções probabilísticas
 - O valor da chave de acesso gera um endereço tão único quanto possível
 - Quando houver coincidência esta situação é chamada de colisão
 - Duas chaves diferentes gerando o mesmo endereço
 - Objetivo das funções probabilísticas
 - Preservar a ordem dos registros
 - Aumentar o grau de unicidade (uniformidade) dos registros sobre o arquivo

Funções para cálculo do endereço

- Exemplo 1: arquivo de empregados
 - Dados os números das matrículas estejam entre 900 e 3150
 - Dados os endereços disponíveis da mídia estejam entre 1 e 37
 - Uma função escolhida para gerar estes endereços pode ser:

$$\text{Função: } E(c) = \frac{(chave - menor\ matrícula) + 1}{(maior\ matrícula - menor\ matrícula) / 37}$$

- Se as chaves de acesso forem 1000, 1400 e 1600
 - Teremos os endereços 2, 9 e 12

Funções para cálculo do endereço

- Exemplo 1: arquivo de empregados

Função: $E(c) = \frac{(chave - 900) + 1}{(3150 - 900) / 37}$

- $E(1000) = 2$
- $E(1400) = 9$
- $E(1600) = 12$

Endereço	Matrícula	Nome	Depto	Salário
1	900	Ademar	A	5000
2	1000	Roberto	B	7500
3	1010	Gerson	A	6000
4	1100	Ieda	C	6000
5				
6	1200	Sandra	C	7500
7	1300	Flavia	C	9000
8				
9	1400	Tatiana	A	8500
10	1480	Maria	B	6500
11				
12	1600	Diogo	B	4500
...

Funções para cálculo do endereço

- Exemplo 2: arquivo de empregados
 - Função que não preserva a ordem dos registros
 - Chamadas de função de aleatorização

Função: $E(c) = (chave \% 31) + 1$

Ordem crescente	↓	Chave	Endereço	Ordem aleatória
		1000	9	
		1050	25	
		1075	22	
		1100	16	
		1300	30	

Tratamentos de colisão

- Tratamento por endereçamento aberto
 - O endereço colidido é guardado no primeiro endereço livre
- Tratamento por encadeamento
 - Busca-se um endereço e adiciona uma ligação ao registro anterior
 - Neste caso, duas alternativas podem ser adotadas
 - Encadeamento puro: os registros que colidem formam uma lista encadeada na área de dados
 - Uso de áreas de extensão: semelhante à abordagem utilizadas em arquivo sequencial indexado

Operações em arquivo direto

- Operações de leitura
 - Acesso sequencial
 - Só é possível quando usada uma função que preserve a ordem dos registros
 - Neste caso, para o acesso sequencial basta ler a área de dados
 - Acesso aleatório
 - Aplica-se a função *hashing*
 - Leitura exaustiva
 - Mesmo princípio do acesso sequencial

Operações em arquivo direto

- Operações de gravação
 - Inserção
 - Aplica-se a função *hashing*
 - Exclusão
 - Usa-se o atributo adicional para exclusão lógica
 - Alteração
 - Quando não há chave de acesso, o registro deve ser localizado e alterado
 - Caso contrário, o registro é excluído e inserido

Arquivo invertido

- Motivação
 - Todas as técnicas de organização de arquivos vistas até então fazem uso da chave primária
 - Porém, existem técnicas voltadas para chaves secundárias
 - Oferece mais eficiência e flexibilidade para o acesso aleatório
 - Cada valor da chave de acesso está ligada uma lista de identificação
 - Esta lista de identificação de registros é chamada de lista invertida
- Estrutura de um arquivo invertido
 - Inversão: conjunto de listas invertidas ligadas a uma chave de acesso
 - Um arquivo pode ter uma ou mais inversões

Estrutura de um arquivo invertido

- Exemplo: arquivo com inversão ligada ao atributo departamento

Endereço	Matrícula	Nome	Data nasc	Depto	Salário
301	1000	Ademar	11/02/1990	A	5000
302	1010	Roberto	17/01/1985	B	7500
303	1020	Gerson	05/12/1988	A	6000
304	1030	Ieda	18/05/1963	C	9000
305	1040	Bernardo	14/12/1992	C	4500
306	1050	Angela	15/02/1995	C	6500

Área de dados

Depto	Endereço
A	301, 303
B	302
C	304, 305, 306

Índice invertido

Arquivo invertido

- Vantagem
 - Permite o acesso direto a um conjunto de registros
- Desvantagem
 - As listas só são válidas para aquela disposição física
 - Se o arquivo for reorganizado, as inversões terão que ser regeneradas
 - Para lidar com isto, implementa-se as listas por chaves primárias
 - Entretanto, há uma perda de eficiência

Organização de arquivos

Projeto de arquivos
Prof. Allan Rodrigo Leite