PRA – Projeto de Arquivos

Organização de arquivos

Prof. Allan Rodrigo Leite

Organização de arquivos

- Armazenamento de pequeno volume de dados
  - Distribuição simples dos registros em um arquivo
  - É eficiente pois a frequência de acessos aleatórios não deve ser elevada

- Armazenamento de grande volume de dados ou aumento da complexidade dos acessos
  - Baixa eficiência no armazenamento dos arquivos e acessos aos registros
  - Requer técnicas sofisticadas de armazenamento e recuperação de dados

Estratégias de organização de arquivos

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

Algumas definições

- Arquivo
  - Coleção de registros lógicos, cada um representando um objeto ou entidade

- Registro lógico (ou apenas registro)
  - Sequência de itens que representam campos ou atributos do registro
  - Um atributo é uma característica ou propriedade constituída por nome, tipo e comprimento
  - Observação: o comprimento pode ser constante ou variável

- Registro físico
  - Armazenamento do arquivo em bloco de registros lógicos
  - Tamanho do bloco coincide com uma unidade de armazenamento do meio físico (por exemplo, setores e trilhas de um hard-disk)
  - Registro físico possibilita a leitura e gravação de registros lógicos
  - Cada bloco armazena um número inteiro de registros

Algumas definições

- 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), mas que pode ter seu valor repetido em diferentes registros
- Chave de acesso
  - Chave utilizada para identificar os registros desejados em uma operação de acesso a um arquivo

Algumas definições

- Argumento de pesquisa
  - Valor da chave de acesso em uma operação
- 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
  - Registros são distribuídos em uma ordem arbitrária, um após o outro, dentro do bloco
  - Em geral a ordem pode ser a mesma da geração dos registros
- Vantagem
  - Simplicidade
- Desvantagem
  - Busca de registro através de acesso sequencial

Arquivo sequencial ordenado

- Os registros estão dispostos ordenadamente
  - Obedece à sequência definida por uma chave primária (chave de ordenação)

Localizar empregado com **matrícula 1030**

Chave de pesquisa: 1030

Chave de ordenação

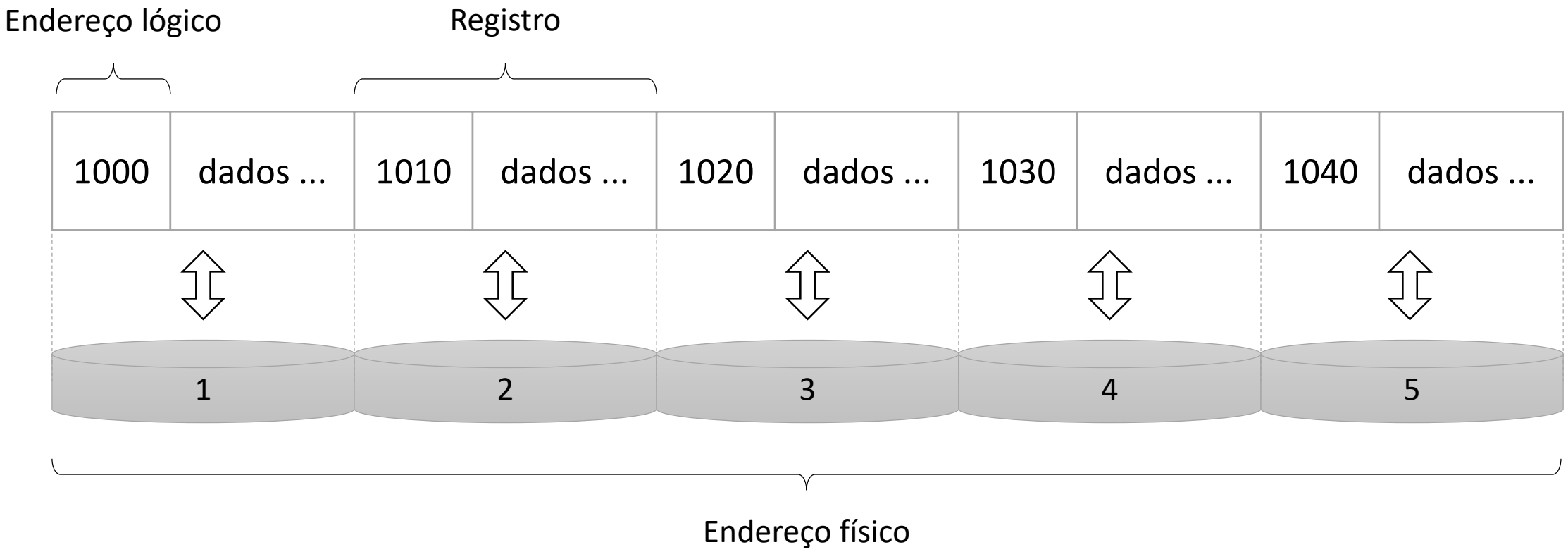
Arquivo: empregado

Matrícula	Nome	Data nascimento	Salário
1000	Ademar	11/02/1990	5000
1010	Roberto	17/01/1985	7500
1020	Gerson	05/12/1988	6000
1030	Ieda	18/05/1963	9000
1040	Bernardo	14/12/1992	4500

Atributos

Arquivo sequencial ordenado

- Estrutura do registro lógico e físico



Arquivo sequencial ordenado

- Principais características
  - Os registros são gravados em ordem sequencial por suas respectivas chaves em uma organização perfeita, tanto lógica quanto física
  - Os registros possuem o mesmo formato, assim cada valor de atributo está associado ao nome do atributo pela sua posição relativa no registro
    - A estrutura (layout) do registro é externa aos dados que ela descreve
    - Esta descrição é declarada nos programas por meio de declarações de tipos e tamanhos
  - Como o formato é único para todas as ocorrências do registro, campos alfanuméricos são dimensionados pelo tamanho máximo
    - Portanto, pode ocorrer desperdício de posições de armazenamento

Arquivo sequencial ordenado

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

- Desvantagens
  - Operações de acesso a um registro cuja chave de acesso não coincide com a chave de ordenação
  - Operações de modificação no arquivo pode requerer uma reorganização
    - Inserção, alteração e remoção de registros

Arquivo sequencial ordenado

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

Operações em arquivo sequencial ordenado

- Acesso sequencial a um registro
  - Consiste na obtenção 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: apresentar os 3 primeiros empregados ordenados pela matrícula

Operações em arquivo sequencial ordenado

- Acesso aleatório a um registro
  - Pesquisa baseada em um argumento de pesquisa
  - Sequência de acesso não mantém necessariamente relação com a ordenação física do arquivo

- Cenários
  - Chave de pesquisa não coincide com chave de ordenação
    - Acesso 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 (mais eficiente)

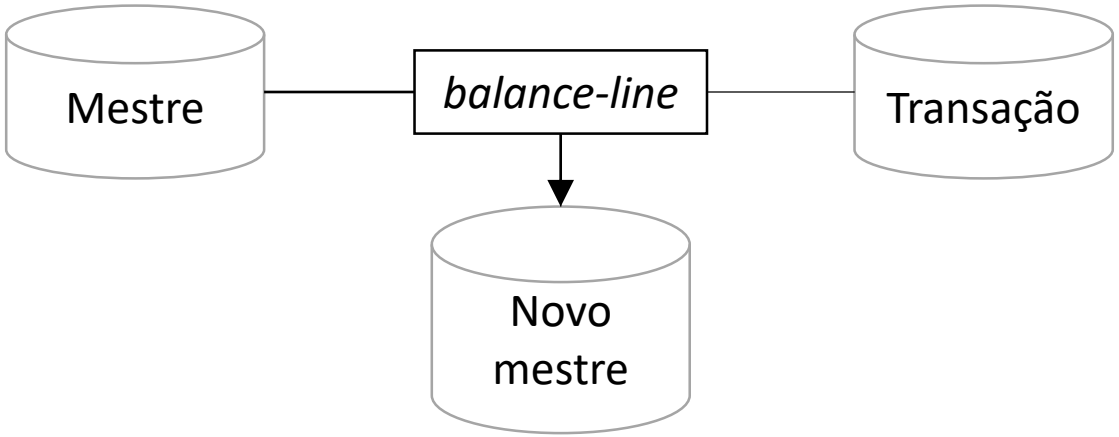
Operações em arquivo sequencial ordenado

- Inserção de um novo registro
  - Utiliza uma técnica conhecida como *balance-line*
  - Inserir um único registro requer o deslocamento dos demais

- *Balance-line*
  - Inserções e alterações são realizadas em um arquivo temporário
  - São inseridos vários registros e somente depois estes serão inseridos no arquivo original
  - Em seguida é realizada a intercalação do arquivo temporário com o arquivo principal, resultando 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 serem incluídos
    - Transação: sequência de operações que 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 campo adicional
    - Campo adicional para indicar o estado do registro como excluído (exclusão lógica)
    - Pesquisa e leitura ignoram os registros marcados como excluídos
- Alteração de um registro existente
  - Usa-se *balance-line*
  - Alteração pode causar aumento do tamanho do registro
  - Alteração pode 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
  - Acessos aleatórios
    - Sequência de acesso nem sempre relação com a ordenação física do arquivo
  - Quando o volume de acessos aleatórios torna-se muito grande, é necessário um estrutura de acesso mais eficiente

- Arquivo sequencial indexado
  - Arquivo sequencial acrescido de índice e área de extensão

Arquivo sequencial indexado

- Um arquivo sequencial indexado é constituído por 3 áreas:
  - Área de índices
    - Arquivo sequencial criado pelo sistema, no qual cada registro estabelece uma divisão na área primária e 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, de forma semelhante a busca por um capítulo de um livro a partir de seu índice
  - Área primária (principal)
    - Reservada para os registros de dados, classificados em ordem ascendente pelo seu campo chave
  - Área de excedentes (*overflow*)
    - Reservada para o acréscimo de novos registros que não podem ser colocados na área principal quando se produz uma inserção no arquivo

Arquivo sequencial indexado

- Índice
  - Formado por uma coleção de pares, onde cada entrada está associada a um valor de chave a 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

Início do bloco

Cada entrada identifica um bloco

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

Arquivo sequencial indexado

- Índice pode ser organizado em múltiplos níveis

Matricula	Endereço
300	1
600	4
*	7

#	Matricula	Endereço
1	100	1
2	200	2
3	300	8
4	400	15
5	500	17
6	600	25
7	700	32
8	800	38
9	*	40

#	Matrícula	Nome	Salário
1	10	Carlos	5000
2	50	Roberto	7000
3	90	Angela	6000
4	105	Claudia	9000
5	120	Afonso	8000
6	180	Enio	5000
7	200	Ademar	4000
...	...	...	...

Índice em 2 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

  - Necessária pois não é viável a implementação da operação de inserção de registros do mesmo modo que nos arquivos sequenciais
    - Os registros podem mudar de endereço exigindo 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

- Acesso
  - Acesso sequencial
    - Direto sobre a área de dados e extensão sem usar o índice
  - Acesso aleatório
    - Uso do índice
    - Pode obter o endereço do próprio registro ou de seu bloco
    - Este último caso requer uma busca dentro do bloco e incluir mais acessos à área de extensão
- Leitura exaustiva (*full-scan*)
  - Igual ao acesso serial

Operações em arquivo sequencial indexado

- Inclusão
  - Usa as áreas de extensão
- Exclusão
  - Pode ser colocada uma marca de excluído no campo situação do registro
- Alteração
  - Pesquisa-se o registro no arquivo
  - Se a alteração não envolver a chave de ordenação, o registro é sobrescrito
  - Se envolver a chave de ordenação, usa-se as operações de exclusão e inclusão

Operações em arquivo sequencial indexado

- Reorganização
  - Desempenho da operações é degradado à medida que ocorre um grande número de inclusões e exclusões
    - Requer periodicamente a exclusão de forma física e lógica os registros excluídos e sanear a área de extensão
    - Neste caso, um novo índice deve ser gerado
  - O limite de reorganização deve ser estabelecido
    - Exemplo, 75% de utilização da área de extensão em horário que o arquivo não tem uso

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
  - Permite com certa facilidade as inserções e exclusões por meio do uso de áreas de extensão

Arquivo indexado

- Motivação
  - Para oferecer um acesso serial eficiente, os arquivos sequenciais ordenados requerem que os registros fisicamente ordenados
  - Isto dificulta a inserção de um registro e exige a utilização de áreas de extensão e da efetivação de reorganizações periódicas
  - Quando a frequência de acessos seriais é baixa e a frequência de acessos aleatórios é alta, a manutenção do sequenciamento torna-se inviável

- 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

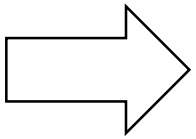
- Suporte a múltiplos índices
  - Podem existir tantos índices quantas forem as chaves de acesso aos registros
  - Um índice contém um conjunto de entradas ordenadas pela chaves de acesso
  - Permite uma busca mais eficiente e o acesso serial ao arquivo
  - Cada entrada do índice contém o valor do atributo e um ponteiro ao endereço do registro que o contém

- Classificações
  - 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

Índice exaustivo (primário)



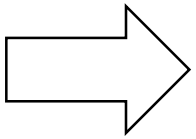
Endereço	Matrícula	Nome	Data nasc	Depto	Salário
301	1000	Ademar	11/02/1990	A	5000
302	1010	Roberto	17/01/1985	B	7500
303	1020	Gerson	05/12/1988	A	6000
304	1030	Ieda	18/05/1963	C	9000
305	1040	Bernardo	14/12/1992	C	4500
306	1050	Angela	15/02/1995	C	6500

Área de dados

Índice seletivo

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

Índice exaustivo (primário)



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

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)

Arquivo indexado

- Acesso
  - Acesso serial
    - Utiliza-se o índice apropriado cuja identificação é simplificada, pois as entradas dos índices são ordenadas
    - Neste caso, a memória mantem um bloco do índice, reduzindo o número de leituras ao disco (memória secundária)
  - Acesso aleatório
    - Requer uma busca sobre o índice
  - Leitura exaustiva (*full-scan*)
    - Para uma leitura exaustiva são realizados sucessivos acessos seriais sobre índices exaustivos

Operações em arquivos indexados

- 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
  - É liberada a área de dados ocupada e são removidas as entradas correspondentes a este 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 via índices

- Desvantagens
  - Acesso serial ineficiente
  - Necessidade de manutenção de um ou mais índices
  - Inserções ou alterações envolvendo atributos associados aos índices

Arquivo direto

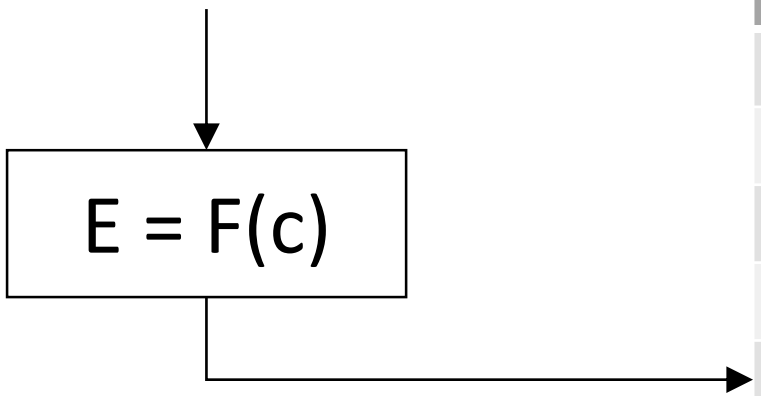
- Motivação
  - Acesso rápido aos registros especificados por argumentos de pesquisa, sem percorrer uma estrutura auxiliar (índice)

- 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
  - Para cada valor da chave de acesso, gera um endereço tão único quanto possível
  - Quando houver coincidência esta situação é chamada de colisão
    - Duas chaves 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
  - Dados os números das matrículas dos empregados esteja entre 900 e 3150
  - Dados os endereços disponíveis estejam entre 1 e 37
  - Uma função escolhida para gerar estes endereços pode ser:

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

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

Funções para cálculo do endereço

- Exemplo 1

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

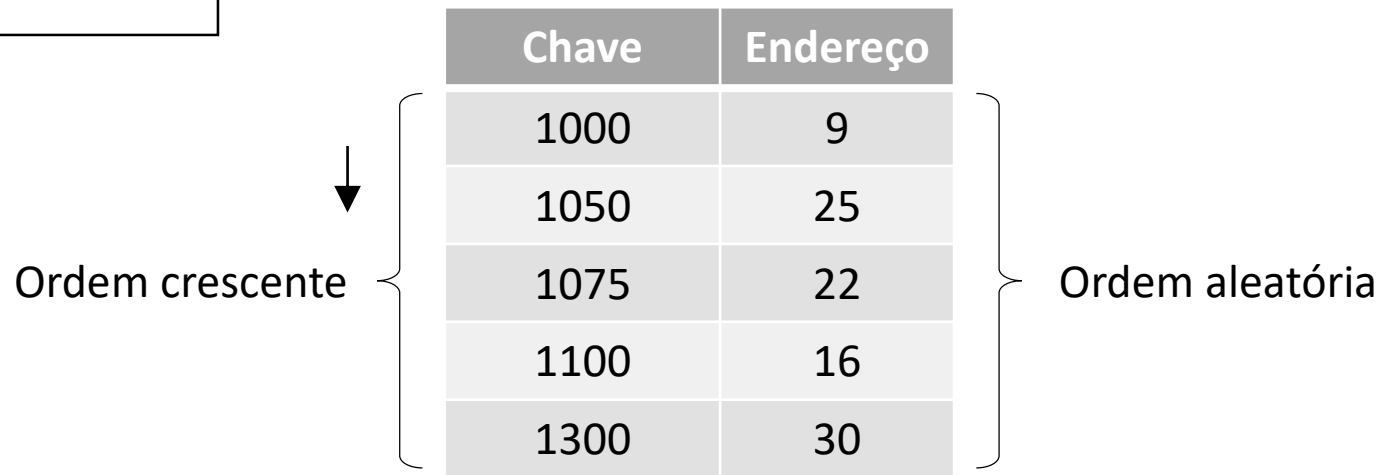
- E(**1000**) = **2**
- E(**1400**) = **9**
- E(**1600**) = **12**

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
  - Funções que não preservam a ordem dos registros
  - Chamadas de funções de aleatorização

Função: E(c) = (*chave* % 31) + 1



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 para encadeá-lo ao anterior
  - Neste caso, duas alternativas podem ser adotadas:
    - Encadeamento puro: registros que colidem formam uma lista encadeada na área de dados
    - Uso de áreas de extensão: semelhante ao usado no arquivo sequencial indexado

Operações em arquivo direto

- Acesso serial
  - Somente é possível se for utilizada uma função que preserve a ordem dos registros
  - Neste caso, para o acesso serial basta ler a área de dados
  - Leitura exaustiva segue o mesmo princípio

- Acesso aleatório e inserção
  - Aplicar a função de cálculo

- Exclusão
  - É acessado o registro e colocada a marca de excluído

- 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
  - Entretanto, existem outras técnicas voltadas para chaves secundárias, com o propósito de resolver o problema de um conjunto de registros
  - Cada valor da chave de acesso está associada uma lista de identificação de registros, chamada lista invertida

- Estrutura de um arquivo invertido
  - Inversão: é o conjunto de listas invertidas associadas a uma chave de acesso
  - Um arquivo pode ter uma ou mais inversões

Estrutura de um arquivo invertido

- Exemplo: arquivo com inversão associada 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

Estrutura de um 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 sofrer uma reorganização, as inversões terão que ser regeradas

  - Para superar esta desvantagem, implementa-se as listas por chaves primárias
    - Entretanto, há uma perda de eficiência

PRA – Projeto de Arquivos

Organização de arquivos

Prof. Allan Rodrigo Leite