Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação
Disciplina de Organização de Arquivos (SCC0215)
docente
Profa. Dra. Cristina Dutra de Aguiar Ciferri (cdac@icmc.usp.br)
alunos PAE
Turma A. Raul Donaire (raul.oliveira@usp.br)
Turma B. João Paulo Clarindo (jpcsantos@usp.br)
monitores
Turma A. Vinicius Ricardo Carvalho (vinicius_carvalho@usp.br)
(telegram: @viniciusRC)
Turma B. Mateus Prado Santos (mateus.prado@usp.br)
colaborador
Matheus Carvalho Raimundo (mcarvalhor@usp.br)

Terceiro Trabalho Prático

Este trabalho tem como objetivo indexar arquivos de dados usando um índice árvore-B.

O trabalho deve ser feito por, no máximo, 2 alunos da mesma turma. Os alunos devem ser o mesmo do primeiro e do segundo trabalhos práticos. Caso haja mudanças, elas devem ser informadas para a docente, os alunos PAE e os monitores. A solução deve ser proposta exclusivamente pelo(s) aluno(s) com base nos conhecimentos adquiridos nas aulas. Consulte as notas de aula e o livro texto quando necessário.

Descrição do arquivo de índice árvore-B

O índice árvore-B com ordem *m* é definido formalmente como descrito a seguir.

1. Cada página (ou nó) do índice árvore-B deve ser, pelo menos, da seguinte forma:

   $< <C_1,P_{R1}>, <C_2,P_{R2}>, ..., <C_{q-1},P_{Rq-1}>, P_1, P_2, ..., P_{q-1}, P_q>$, onde $(q \leq m)$ e

   - Cada C_i $(1 \leq i \leq q-1)$ é uma chave de busca.
   - Cada P_{Ri} $(1 \leq i \leq q-1)$ é um campo de referência para o registro no arquivo de dados que contém o registro de dados correspondente a C_i.
   - Cada P_j $(1 \leq j \leq q)$ é um campo de referência para uma subárvore ou assume o valor -1 caso não exista subárvore (ou seja, caso seja um nó folha).

2. Dentro de cada página (ou seja, as chaves de busca são ordenadas)
   - $C_1 < C_2 < ... < C_{q-1}$.
3. Para todos os valores X da chave na subárvore apontada por P_i:
   - $C_{i-1} < X < C_i$ para $1 < i < q$
   - $X < C_i$ para $i = 1$
   - $C_{i-1} < X$ para i = q.
4. Cada página possui um máximo de m descendentes.
5. Cada página, exceto a raiz e as folhas, possui no mínimo $\lceil m/2 \rceil$ descendentes (*taxa de ocupação*).
6. A raiz possui pelo menos 2 descendentes, a menos que seja um nó folha.
7. Todas as folhas aparecem no mesmo nível.
8. Uma página não folha com k descendentes possui k-1 chaves.
9. Uma página folha possui no mínimo $\lceil m/2 \rceil - 1$ chaves e no máximo $m - 1$ chaves (*taxa de ocupação*).

Descrição do Registro de Cabeçalho. O registro de cabeçalho deve conter os seguintes campos:

- *status*: indica a consistência do arquivo de índice, devido à queda de energia, travamento do programa, etc. Pode assumir os valores 0, para indicar que o arquivo de dados está inconsistente, ou 1, para indicar que o arquivo de dados está consistente. Ao se abrir um arquivo para escrita, seu *status* deve ser 0 e, ao finalizar o uso desse arquivo, seu *status* deve ser 1 – tamanho: *string* de 1 byte.
- *noRaiz*: armazena o RRN do nó (página) raiz do índice árvore-B. Quando a árvore-B está vazia, noRaiz = -1 – tamanho: inteiro de 4 *bytes*
- *nroNiveis:* armazena o número de níveis do índice árvore-B. Inicialmente, a árvore-B está vazia e, portanto, *nroNiveis* = 0. Quando as primeiras *m* chaves de busca são inseridas, o nó raiz é igual ao nó folha e, nesse caso, *nroNiveis* = 1. Quando ocorrer o primeiro *split*, *nroNiveis* = 2. Depois disso, cada vez que o

nível da árvore aumentar, *nroNiveis* deve ser incrementado – tamanho: inteiro de 4 *bytes*

- *proxRRN:* armazena o valor do próximo RRN a ser usado para conter um nó (página da árvore-B). Inicialmente, a árvore-B está vazia e, portanto, *proxRRN* = 0. Quando o primeiro nó é criado (nó folha = nó raiz), *proxRRN* = 1. Depois, quando primeiro *split* acontece, *proxRRN* = 3. A cada nó criado da árvore-B, *proxRRN* é incrementado – tamanho: inteiro de 4 *bytes*
- *nroChaves:* armazena o número de chaves de busca indexadas no índice árvore-B. Inicialmente, a árvore-B está vazia e, portanto, *nroChaves* = 0. A cada chave de busca inserida na árvore-B, *nroChaves* é incrementado – tamanho: inteiro de 4 *bytes*

Representação Gráfica do Registro de Cabeçalho. O tamanho do registro de cabeçalho é de 72 bytes, representado da seguinte forma:

1 byte	4 bytes				4 bytes				4 bytes				4 bytes				55 bytes	
status	*noRaiz*				*nroNiveis*				*proxRRN*				*nroChaves*				*lixo (caractere '$')*	
0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*...*	*71*

Observações Importantes.

- O registro de cabeçalho deve seguir estritamente a ordem definida na sua representação gráfica.
- Os nomes dos atributos também devem seguir estritamente os nomes definidos na especificação dos mesmos.
- Para seguir a especificação do conceito de árvore-B, o nó da árvore-B deve obrigatoriamente ser do tamanho de uma página de disco. Entretanto, isso não será seguido neste trabalho para simplificar a quantidade de chaves de busca que são armazenadas no nó. Lembrando também que as páginas de disco têm potência de 2, o que também não será seguido neste trabalho por simplificação.
- Os 59 bytes restantes devem ser preenchidos com lixo. O lixo é representado pelo caractere '$'.

Descrição do Registro de Dados. Deve ser considerada a seguinte organização: campos de tamanho fixo e registros de tamanho fixo. Em adição ao Item 1 da definição formal do índice árvore-B, cada nó (página) da árvore também deve armazenar dois outros campos:

- *nível*, indicando o nível no qual o nó se encontra. Quando um nó é um nó-folha, *nível* = 1. Quando um nó aponta para um nó-folha, *nível* = 2. Quando um nó aponta para um outro nó que, por sua vez, aponta para um nó-folha, *nível* = 3. E assim sucessivamente – tamanho: inteiro de 4 *bytes*
- *n*, indicando o número de chaves presentes no nó – tamanho: inteiro de 4 *bytes*.

A ordem da árvore-B é 6, ou seja, $m = 6$. Portanto, um nó (página) terá 5 chaves e 6 descendentes. A chave de busca é o campo *idNascimento*.

Considere que encontram-se implementadas as rotinas de *split* (para inserção) e de concatenação (para remoção). Considere que a rotina de redistribuição durante a inserção não encontra-se implementada. Considere que a distribuição das chaves de busca deve ser o mais uniforme possível. Considere também que a chave de busca a ser promovida deve ser a primeira chave do novo nó resultante do particionamento (ou seja, o primeiro elemento do segundo nó é a chave promovida durante o particionamento). Quando necessário, o nó mais à esquerda deverá conter uma chave de busca a mais. Considere que a página sendo criada é sempre a página à direita.

Representação Gráfica de um Nó (Página/Registro de Dados) do índice. O tamanho do registro de cabeçalho é de 72 bytes, representado da seguinte forma:

4 bytes				4 bytes				4 bytes	4 bytes		4 bytes	4 bytes	4 bytes	4 bytes	4 bytes				
nível				*n*				C_1	P_{R1}	...	C_5	P_{R5}	P_1	...	P_6				
0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	...							*68*	*69*	*70*	*71*	

Observações Importantes.

- Cada registro de dados deve seguir estritamente a ordem definida na sua representação gráfica.

- Os nomes dos atributos também devem seguir estritamente os nomes definidos na especificação dos mesmos.
- Quando um nó (página) do índice tiver chaves de busca que não forem preenchidas, deve ser armazenado o caractere '$'. O valor -1 deve ser usado para denotar que um ponteiro P_i ($1 \leq i \leq m$) de um nó é nulo.

Programa

Descrição Geral. possa realizar operações sobre um índice árvore-B e que possa *inserir* e *buscar* dados de um arquivo de dados que está indexado por esse índice.

Importante. A definição da sintaxe de cada comando bem como sua saída devem seguir estritamente as especificações definidas em cada funcionalidade. Para especificar a sintaxe de execução, considere que o programa seja chamado de "programaTrab". Essas orientações devem ser seguidas uma vez que a correção do funcionamento do programa se dará de forma automática. De forma geral, a primeira entrada da entrada padrão é sempre o identificador de suas funcionalidades, conforme especificado a seguir.

Descrição Específica. O programa deve oferecer as seguintes funcionalidades:

[8] Crie um arquivo de índice árvore-B para um arquivo de dados de entrada já existente, que é o arquivo de dados definido de acordo com a especificação do primeiro trabalho prático, e que pode conter registros logicamente removidos. O campo a ser indexado é *idNascimento*. Registros logicamente removidos presentes no arquivo de dados de entrada não devem ter suas chaves de busca correspondentes no arquivo de índice. A inserção no arquivo de índice deve ser feita um-a-um. Ou seja, para cada registro não removido presente no arquivo de dados, deve ser feita a inserção de sua chave de busca correspondente no arquivo de índice árvore-B. A

manipulação do arquivo de índice árvore-B deve ser feita em disco, de acordo com o conteúdo ministrado em sala de aula. Antes de terminar a execução da funcionalidade, deve ser utilizada a função binarioNaTela, disponibilizada na página do projeto da disciplina, para mostrar a saída do arquivo de índice árvore-B.

Entrada do programa para a funcionalidade [8]:

```
8 arquivoEntrada.bin arquivoIndiceIdNascimento.bin
```

onde:

- `arquivoEntrada.bin` é um arquivo binário de entrada que segue as mesmas especificações do primeiro trabalho prático, e que contém dados desordenados e registros logicamente removidos.
- `arquivoIndiceIdNascimento.bin` é o arquivo binário de índice árvore-B que indexa o campo *idNascimento*. Esse arquivo deve seguir as especificações definidas neste trabalho prático.

Saída caso o programa seja executado com sucesso:

```
Listar o arquivo binário arquivoIndiceIdNascimento.bin usando a função binarioNaTela.
```

Mensagem de saída caso algum erro seja encontrado:

```
Falha no processamento do arquivo.
```

Exemplo de execução:

```
./programaTrab
8 arquivoEntrada.bin arquivoIndiceIdNascimento.bin
usar a função binarioNaTela antes de terminar a execução da funcionalidade, para mostrar a saída do arquivo arquivoIndiceIdNascimento.bin, o qual indexa o campo idNascimento.
```

[9] Permita a recuperação dos dados de todos os registros que satisfaçam um critério de busca determinado pelo usuário sobre o campo *idNascimento*, usando o índice árvore-B criado na funcionalidade [8]. Note que somente o campo *idNascimento* deve ser utilizado como forma de busca, desde que o índice criado na funcionalidade [8] indexa chaves de busca desse campo. Essa funcionalidade pode retornar 0 registros (quando nenhum satisfaz ao critério de busca) ou 1 registro (desde que o identificador do servidor não aceita valores repetidos). A manipulação do arquivo de índice árvore-B deve ser feita em disco, de acordo com o conteúdo ministrado em sala de aula. Os dados devem ser mostrados no mesmo formato definido para a funcionalidade [2]. Registros marcados como logicamente removidos não devem ser exibidos. Depois de mostrar todos os registros, deve ser mostrado na saída padrão a quantidade de páginas da árvore-B que foram acessadas. Lembre-se que cada nó da árvore-B é armazenado em uma página de disco diferente. Nesse cálculo, considere que o registro de cabeçalho da árvore-B encontra-se em RAM e que, portanto, não deve ser considerado nos cálculos.

Sintaxe do comando para a funcionalidade [9]:

```
9 arquivoEntrada.bin arquivoIndiceIdNascimento.bin idNascimento valor
```

onde:

- `arquivoEntrada.bin` é um arquivo binário de entrada que segue as mesmas especificações do primeiro trabalho prático, e que contém dados desordenados e registros logicamente removidos.
- `arquivoIndiceIdNascimento.bin` é o arquivo binário de índice árvore-B que indexa o campo *idNascimento*. Esse arquivo deve seguir as especificações definidas neste trabalho prático.
- `idNascimento` é o nome do campo que representa a chave de busca.
- `valor` é o número que representa o `idNascimento` sendo procurado.

Saída caso o programa seja executado com sucesso:

Apenas alguns dados dos registros de dados devem ser exibidos na saída padrão, a saber: "Nasceu em " cidadeBebe "/" estadoBebe ", em " dataNascimento ", um bebe de sexo " sexoBebe "." O dado referente ao campo sexoBebe deve ser escrito da seguinte forma: "IGNORADO" caso sexoBebe = '0', "MASCULINO" caso sexoBebe = '1' e "FEMININO" caso sexoBebe = '2'. Campos que tiverem o valor nulo não devem ser mostrados, sendo substituídos pelo caractere "-".

Mensagem de saída caso não seja encontrado o registro que contém o valor do campo ou o campo pertence a um registro que esteja removido:

```
Registro inexistente.
```

Mensagem de saída caso algum erro seja encontrado:

```
Falha no processamento do arquivo.
```

Exemplo de execução (é mostrado um registro ilustrativo):

```
./programaTrab
9 arquivoEntrada.bin arquivoIndiceIdNascimento.bin idNascimento 3
Nasceu em ARARAQUARA/SP, em -, um bebe do sexo IGNORADO.
Quantidade de paginas da arvore-B acessadas: 2
```

[10] Estenda a funcionalidade [6] de forma que, a cada inserção de um registro no arquivo de dados, a chave de busca correspondente a essa inserção seja inserida no arquivo de índice árvore-B. Antes de terminar a execução da funcionalidade, deve ser utilizada a função binarioNaTela, disponibilizada na página do projeto da disciplina, para mostrar a saída do arquivo de índice árvore-B.

Relembrando a funcionalidade [6], ela especifica o seguinte. Permita a inserção de registros adicionais, baseado na *abordagem estática* de registros logicamente removidos. A implementação dessa funcionalidade deve seguir estritamente a matéria apresentada em sala de aula. Não é necessário realizar o tratamento de truncamento de dados. Portanto, a soma dos tamanhos para os campos de tamanho variável nunca deve ultrapassar o tamanho do registro de tamanho fixo. Campos com valores nulos, na entrada da funcionalidade, devem ser identificados com NULO. A funcionalidade [6] deve ser executada *n* vezes seguidas. Lembre-se de manipular o campo "status" do registro de cabeçalho adequadamente. Quando o arquivo começar a ser modificado, deve ser marcado como inconsistente. Ao final de todas as modificações, deve ser marcado como consistente.

Entrada do programa para a funcionalidade [10] (mesma entrada da [6]):

`10 arquivoEntrada.bin arquivoIndiceIdNascimento.bin n`
`cidadeMae`$_1$ `cidadeBebe`$_1$ `idNascimento`$_1$ `idadeMae`$_1$ `dataNascimento`$_1$ `sexoBebe`$_1$ `estadoMae`$_1$ `estadoBebe`$_1$
`...`
`cidadeMae`$_n$ `cidadeBebe`$_n$ `idNascimento`$_n$ `idadeMae`$_n$ `dataNascimento`$_n$ `sexoBebe`$_n$ `estadoMae`$_n$ `estadoBebe`$_n$

onde:

- `arquivoEntrada.bin` é um arquivo binário de entrada que segue as mesmas especificações do primeiro trabalho prático, e que contém dados desordenados e registros logicamente removidos.
- `arquivoIndiceIdNascimento.bin` é o arquivo binário de índice árvore-B que indexa o campo *idNascimento*. Esse arquivo deve seguir as especificações definidas neste trabalho prático.
- `n` é o número de inserções a serem realizadas. Para cada inserção, deve ser informado os valores a serem inseridos no arquivo, para os campos especificados na mesma ordem que a definida nesse trabalho prático, a saber: `cidadeMae`, `cidadeBebe`, `idNascimento`, `idadeMae`, `dataNascimento`, `sexoBebe`, `estadoMae`, `estadoBebe`. Não existe truncamento de dados. Valores nulos devem ser identificados, na entrada da funcionalidade, por NULO. Cada uma das `n` inserções deve ser especificada em uma linha diferente. Deve ser deixado um espaço em branco entre os valores dos campos. Os valores dos campos do tipo *string* devem ser especificados entre aspas duplas (").

Saída caso o programa seja executado com sucesso:

`Listar o arquivo binário arquivoIndiceIdNascimento.bin usando a função binarioNaTela.`

Mensagem de saída caso algum erro seja encontrado:

`Falha no processamento do arquivo.`

Exemplo de execução:

```
./programaTrab
10 arquivoEntrada.bin arquivoIndiceIdNascimento.bin
“MATAO“ “RIBEIRAO PRETO“ 3 28 “2019-05-20“ “2“ “SP” “SP”
“ARARAQUARA“ “ARARAQUARA“ 5 NULO NULO “1“ “SP” “SP”
```

`usar a função binarioNaTela antes de terminar a execução da funcionalidade, para mostrar a saída do arquivo arquivoIndiceIdNascimento.bin, o qual indexa o campo idNascimento.`

Restrições

As seguintes restrições têm que ser garantidas no desenvolvimento do trabalho.

[1] O arquivo de dados deve ser gravado em disco no **modo binário**. O modo texto não pode ser usado.

[2] Os dados do registro descrevem os nomes dos campos, os quais não podem ser alterados. Ademais, todos os campos devem estar presentes na implementação, e nenhum campo adicional pode ser incluído. O tamanho e a ordem de cada campo deve obrigatoriamente seguir a especificação.

[3] Deve haver a manipulação de valores nulos, conforme as instruções definidas.

[4] Não é necessário realizar o tratamento de truncamento de dados.

[5] Devem ser exibidos avisos ou mensagens de erro de acordo com a especificação de cada funcionalidade.

[6] Os dados devem ser obrigatoriamente escritos campo a campo. Ou seja, não é possível escrever os dados registro a registro. Essa restrição refere-se à entrada/saída, ou seja, à forma como os dados são escritos no arquivo. Para se fazer a busca, é possível caminhar no arquivo registro a registro, já que se sabe o tamanho do registro.

[7] O(s) aluno(s) que desenvolveu(desenvolveram) o trabalho prático deve(m) constar como comentário no início do código (i.e. NUSP e nome do aluno). Para trabalhos desenvolvidos por mais do que um aluno, não será atribuída nota ao aluno cujos dados não constarem no código fonte.

[8] Todo código fonte deve ser documentado. A **documentação interna** inclui, dentre outros, a documentação de procedimentos, de funções, de variáveis, de partes do

código fonte que realizam tarefas específicas. Ou seja, o código fonte deve ser documentado tanto em nível de rotinas quanto em nível de variáveis e blocos funcionais.

[9] A implementação deve ser realizada usando a linguagem de programação C. As funções das bibliotecas <stdio.h> devem ser utilizadas para operações relacionadas à escrita e leitura dos arquivos. A implementação não pode ser feita em qualquer outra linguagem de programação. O programa executará no [run.codes].

Fundamentação Teórica

Conceitos e características dos diversos métodos para representar os conceitos de campo e de registro em um arquivo de dados podem ser encontrados nos *slides* de sala de aula e também nas páginas 96 a 107 do livro *File Structures* (*second edition*), de Michael J. Folk e Bill Zoellick.

Material para Entregar

Arquivo compactado. Deve ser preparado um arquivo .zip contendo:

- Código fonte do programa devidamente documentado.
- Makefile para a compilação do programa.

Instruções para fazer o arquivo makefile. No [run.codes] tem uma orientação para que, no makefile, a diretiva "all" contenha apenas o comando para compilar seu programa e, na diretiva "run", apenas o comando para executá-lo. Assim, a forma mais simples de se fazer o arquivo makefile é:

```
all:
	gcc -o programaTrab *.c
run:
	./programaTrab
```

Lembrando que *.c já engloba todos os arquivos .c presentes no seu zip. Adicionalmente, no arquivo Makefile é importante se ter um *tab* nos locais colocados acima, senão ele pode não funcionar.

Instruções de entrega. A entrega deve ser feita via [run.codes]:

- página: https://run.codes/Users/login
- **Turma A** (segunda-feira): código de matrícula: **SBRZ**
- **Turma B** (terça-feira): código de matrícula: **24AK**

Critério de Correção

Critério de avaliação do trabalho. Na correção do trabalho, serão ponderados os seguintes aspectos.

- Corretude da execução do programa.
- Atendimento às especificações do registro de cabeçalho e dos registros de dados.
- Atendimento às especificações da sintaxe dos comandos de cada funcionalidade e do formato de saída da execução de cada funcionalidade.
- Qualidade da documentação entregue. A documentação interna terá um peso considerável no trabalho.

Restrições adicionais sobre o critério de correção.

- A não execução de um programa devido a erros de compilação implica que a nota final da parte do trabalho será igual a zero (0).

- O não atendimento às especificações do registro de cabeçalho e dos registros de dados implica que haverá uma diminuição expressiva na nota do trabalho.
- O não atendimento às especificações de sintaxe dos comandos de cada funcionalidade e do formato de saída da execução de cada funcionalidade implica que haverá uma diminuição expressiva na nota do trabalho.
- A ausência da documentação implica que haverá uma diminuição expressiva na nota do trabalho.
- A inserção de palavras ofensivas nos arquivos e em qualquer outro material entregue implica que a nota final da parte do trabalho será igual a zero (0).
- Em caso de plágio, as notas dos trabalhos envolvidos serão zero (0).

Data de Entrega do Trabalho

Na data especificada na página da disciplina.

Bom Trabalho !