Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação
Disciplina de Estrutura de Dados III (SCC0607)

Docente
Profa. Dra. Cristina Dutra de Aguiar
cdac@icmc.usp.br

Monitores
Eduardo Souza Rocha
eduardos.rocha17@usp.br ou telegram: @edwolt
Beatriz Aimee Teixeira Furtado Braga
beatriztfb@usp.br ou telegram: @bia_aimee
Heitor Tanoue de Mello
heitortanoue@usp.br ou *telegram: @heitortanoue*

Trabalho Introdutório

Este trabalho tem como objetivo obter dados de um arquivo de entrada e gerar um arquivo binário com esses dados, bem como realizar operações de busca. Ele é um trabalho introdutório, de forma que será usado como base para o desenvolvimento de todos os demais trabalhos da disciplina.

O trabalho deve ser feito por 2 alunos. A solução deve ser proposta exclusivamente pelos alunos com base nos conhecimentos adquiridos nas aulas. Consulte as notas de aula e o livro texto quando necessário.

Fundamentos da disciplina de Bases de Dados

A disciplina de Estrutura de Dados III é uma disciplina fundamental para a disciplina de Bases de Dados. A definição dos trabalhos práticos é feita considerando esse aspecto, ou seja, os trabalhos são especificados em termos de várias funcionalidades, e essas funcionalidades são relacionadas tanto com desafios enfrentados no mercado de trabalho quanto com as funcionalidades da linguagem SQL (*Structured Query Language*), que é a linguagem utilizada por sistemas gerenciadores de banco de dados (SGBDs) relacionais. As características e o detalhamento de SQL serão aprendidos na disciplina de Bases de Dados. Porém, por meio do desenvolvimento deste trabalho prático, os alunos podem entrar em contato com alguns comandos da linguagem SQL e verificar como eles são implementados.

Os trabalhos práticos têm como objetivo armazenar e recuperar dados referentes aos ecossistemas tecnológicos e como as tecnologias estão relacionadas entre si. Esses dados encontram-se disponíveis no Stack Overflow e podem ser obtidos a partir da URL https://www.kaggle.com/datasets/stackoverflow/stack-overflow-tag-network?select=stack_network_nodes.csv.

Descrição do Arquivo de Dados

Um arquivo de dados possui um registro de cabeçalho e 0 ou mais registros de dados. A descrição do registro de cabeçalho é feita conforme a definição a seguir.

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

- *status*: indica a consistência do arquivo de dados, 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.
- *proxRRN*: armazena o valor do próximo RRN (número relativo do registro) disponível. Deve ser iniciado com o valor '0' e deve ser alterado sempre que necessário – tamanho: inteiro de 4 bytes.
- *nroTecnologias*: indica a quantidade de tecnologias diferentes que estão armazenadas no arquivo de dados. Note que, se duas ou mais tecnologias têm o mesmo nome, elas são consideradas a mesma tecnologia. Deve ser iniciado com o valor '0' e deve ser alterado sempre que necessário – tamanho: *inteiro* de 4 bytes.
- *nroParesTecnologias*: indica a quantidade de pares (nomeTecOrigem, nomeTecDestino) diferentes que estão armazenados no arquivo de dados. Deve ser iniciado com o valor '0' e deve ser alterado sempre que necessário – tamanho: *inteiro* de 4 bytes.

Representação Gráfica do Registro de Cabeçalho. O registro de cabeçalho deve ser representado da seguinte forma:

1 byte	4 bytes	4 bytes	4 bytes
status	*proxRRN*	*nroTecnologias*	*nroParesTecnologias*
0	1...4	5...8	9...12

Observações Importantes.

- O registro de cabeçalho deve seguir estritamente a ordem definida na sua representação gráfica.
- Os campos são de tamanho fixo. Portanto, os valores que forem armazenados não devem ser finalizados por '\0'.
- Neste projeto, o conceito de página de disco não está sendo considerado.

Registros de Dados. Os registros de dados são de tamanho fixo, com campos de tamanho fixo e campos de tamanho variável. Para os campos de tamanho variável, deve ser usado o método indicador de tamanho.

Os campos de tamanho fixo são definidos da seguinte forma:

- *grupo*: grupo ao qual a tecnologia de origem pertence – tamanho: *inteiro* de 4 bytes.
- *popularidade*: total de milhões de acessos que existe no Stack Overflow para aquela tecnologia de origem – tamanho: *inteiro* de 4 bytes.
- *peso*: frequência com que as *tags* de tecnologia no Stack Overflow aparecem juntas em relação à frequência com que aparecem separadamente. Uma tag representa a ideia de relacionamentos entre tecnologias – tamanho: *inteiro* de 4 bytes.

Os campos de tamanho variável são definidos da seguinte forma:

- *nomeTecnologiaOrigem*: nome da tecnologia de origem – *string*
- *nomeTecnologiaDestino:* nome da tecnologia de destino – *string*

Adicionalmente, o seguinte campo de tamanho fixo também compõe cada registro.

- *removido*: indica se o registro está logicamente removido. Pode assumir os valores '1', para indicar que o registro está marcado como logicamente removido, ou '0', para indicar que o registro não está marcado como removido. – tamanho: *string* de 1 byte.

Os dados são fornecidos juntamente com a especificação deste trabalho prático por meio de um arquivo .csv, sendo que sua especificação está disponível na página da disciplina. No arquivo .csv, o separador de campos é vírgula (,) e o primeiro registro especifica o que cada coluna representa (ou seja, contém a descrição do conteúdo das colunas).

Representação Gráfica dos Registros de Dados. Cada registro de dados deve ter o tamanho de 76 bytes, representado da seguinte forma:

1 byte	4 bytes	4 bytes	4 bytes	4 bytes	variável	4 bytes	variável
removido	*grupo*	*popularidade*	*peso*	*tamanho Tecnologia Origem*	*nome Tecnologia Origem*	*tamanho Tecnologia Destino*	*nome Tecnologia Origem*
0	1 ... 4	5 ... 8	9 ...12	13 ... 16	17 ...	...	...

Observações Importantes.

- Cada registro de dados deve seguir estritamente a ordem definida na sua representação gráfica.
- As *strings* de tamanho variável não devem ser finalizadas com '\0', desde que é utilizado o método indicador de tamanho.
- Os valores nulos nos campos inteiros de tamanho fixo devem ser devem ser representados pelo valor -1.
- Os valores nulos nos campos de tamanho variável devem ser manipulados da seguinte forma: o indicador de tamanho deve armazenar o valor zero.
- Deve ser feita a diferenciação entre o espaço utilizado e o lixo. Sempre que houver lixo, ele deve ser identificado pelo caractere '$'. Nenhum *byte* do registro deve permanecer vazio, ou seja, cada *byte* deve armazenar um valor válido ou '$'.

- Não existe a necessidade de truncamento dos dados. O arquivo .csv com os dados de entrada já garante essa característica.
- Neste projeto, o conceito de página de disco não está sendo considerado.

Programa

Descrição Geral. Implemente um programa em C por meio do qual o usuário possa obter dados de um arquivo de entrada, gerar arquivos binários com esses dados e realizar operações de leitura e busca.

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.

Modularização. É importante modularizar o código. Trechos de programa que aparecerem várias vezes devem ser modularizados em funções e procedimentos.

Descrição Específica. O programa deve oferecer as seguintes funcionalidades:

Na linguagem SQL, o comando CREATE TABLE é usado para criar uma tabela, a qual é implementada como um arquivo. Geralmente, uma tabela possui um nome (que corresponde ao nome do arquivo) e várias colunas, as quais correspondem aos campos dos registros do arquivo de dados. A funcionalidade [1] representa um exemplo de implementação do comando CREATE TABLE.

[1] Permita a leitura de vários registros obtidos a partir de um arquivo de entrada no formato csv e a gravação desses registros em um arquivo de dados de saída. O arquivo de entrada no formato csv é fornecido juntamente com a especificação do projeto, enquanto o arquivo de dados de saída deve ser gerado de acordo com as especificações deste trabalho prático. 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 binário.

Entrada do programa para a funcionalidade [1]:

```
1 arquivoEntrada.csv arquivoSaida.bin
```

onde:

- `arquivoEntrada.csv` é um arquivo .csv que contém os valores dos campos dos registros a serem armazenados no arquivo binário.
- `arquivoSaida.bin` é o arquivo binário gerado conforme as especificações descritas neste trabalho prático.

Saída caso o programa seja executado com sucesso:

```
Listar o arquivo de saída no formato binário usando a função fornecida binarioNaTela.
```

Mensagem de saída caso algum erro seja encontrado:

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

Exemplo de execução:

```
./programaTrab
1 tecnologias.csv tecnologias.bin
usar a função binarioNaTela antes de terminar a execução da funcionalidade, para mostrar a saída do arquivo tecnologias.bin.
```

Na linguagem SQL, o comando SELECT é usado para listar os dados de uma tabela. Existem várias cláusulas que compõem o comando SELECT. O comando mais básico consiste em especificar as cláusulas SELECT e FROM, da seguinte forma:

SELECT lista de colunas (ou seja, campos a serem exibidos na resposta)
FROM tabela (ou seja, arquivo que contém os campos)

A funcionalidade [2] representa um exemplo de implementação do comando SELECT. Como todos os registros devem ser recuperados nessa funcionalidade, sua implementação consiste em percorrer sequencialmente o arquivo.

[2] Permita a recuperação dos dados de todos os registros armazenados em um arquivo de dados de entrada, mostrando os dados de forma organizada na saída padrão para permitir a distinção dos campos e registros. O tratamento de 'lixo' deve ser feito de forma a permitir a exibição apropriada dos dados. Registros marcados como logicamente removidos não devem ser exibidos.

Entrada do programa para a funcionalidade [2]:

`2 arquivoEntrada.bin`

onde:

- `arquivoEntrada.bin` é o arquivo binário gerado conforme as especificações descritas neste trabalho prático.

Saída caso o programa seja executado com sucesso:

`Para cada registro, os valores de seus campos devem ser mostrados em uma única linha, sendo separados por vírgula. Caso o campo seja nulo, deve ser exibido: NULO. Ordem de exibição dos campos: nomeTecnologiaOrigem, grupo, popularidade, nomeTecnologiaDestino, peso, conforme ilustrado no` **`exemplo de execução.`**

Mensagem de saída caso não existam registros:

`Registro inexistente.`

Mensagem de saída caso algum erro seja encontrado:

`Falha no processamento do arquivo.`

Exemplo de execução (é mostrado um exemplo ilustrativo)**:**

```
./programaTrab
2 tecnologias.bin
AZURE, 2, 14, .NET, 22
SQL-SERVER, 2, 65, .NET, 33
ASP.NET, 2, 130, .NET, 6
...
```

Conforme visto na funcionalidade [2], na linguagem SQL o comando SELECT é usado para listar os dados de uma tabela. Existem várias cláusulas que compõem o comando SELECT. Além das cláusulas SELECT e FROM, outra cláusula muito comum é a cláusula WHERE, que permite que seja definido um critério de busca sobre um ou mais campos, o qual é nomeado como critério de seleção.

SELECT lista de colunas (ou seja, campos a serem exibidos na resposta)
FROM tabela (ou seja, arquivo que contém os campos)
WHERE critério de seleção (ou seja, critério de busca)

A funcionalidade [3] representa um exemplo de implementação do comando SELECT considerando a cláusula WHERE. Como não existe índice definido sobre os campos dos registros, a implementação dessa funcionalidade consiste em percorrer sequencialmente o arquivo.

[3] Permita a recuperação dos dados de todos os registros de um arquivo de dados de entrada, de forma que esses registros satisfaçam um critério de busca determinado pelo usuário. Qualquer campo pode ser utilizado como forma de busca. Adicionalmente, apenas um campo pode ser utilizado na busca. Por exemplo, é possível realizar a busca considerando somente o campo *peso* ou somente o campo *nomeTecnologiaDestino*. Esta funcionalidade pode retornar 0 registros (quando nenhum satisfaz ao critério de busca), 1 registro (quando apenas um satisfaz ao critério de busca), ou vários registros. A funcionalidade [3] deve ser executada *n* vezes seguidas. Em situações nas quais um determinado critério de busca não seja satisfeito, ou seja, caso a solicitação do usuário não retorne nenhum registro a ser buscado, o programa deve continuar a executar as buscas até completar as *n* vezes seguidas. Os valores dos campos do tipo *string* devem ser especificados entre aspas duplas ("). Para a manipulação de *strings* com aspas duplas, pode-se usar a função scan_quote_string disponibilizada na página do projeto da disciplina. Registros marcados como logicamente removidos não devem ser exibidos. O arquivo de dados de entrada deve ser percorrido apropriadamente.

Sintaxe do comando para a funcionalidade [3]:

`3 arquivoEntrada.bin n`
`nomeCampo`$_1$ `valorCampo`$_1$
`nomeCampo`$_2$ `valorCampo`$_2$
`...`
`nomeCampo`$_n$ `valorCampo`$_n$

onde:

- `arquivoEntrada.bin` é o arquivo binário de entrada, o qual foi gerado conforme as especificações descritas neste trabalho prático.

- `n` é a quantidade de vezes que a busca deve ser realizada. Para cada busca, deve ser indicado o `nome do Campo` e o `valor do Campo` utilizados na busca. Cada um dos `n` (`nomeCampo valorCampo`) deve ser especificado em uma linha diferente. Deve ser deixado um espaço em branco entre `nomeCampo` e `valorCampo`. Os valores dos campos do tipo *string* devem ser especificados entre aspas duplas (").

Saída caso o programa seja executado com sucesso:

`Para cada valor de n, exiba a seguinte saída. Para cada registro, os valores de seus campos devem ser mostrados em uma única linha, sendo separados por vírgula. Caso o campo seja nulo, deve ser exibido: NULO. Ordem de exibição dos campos: nomeTecnologiaOrigem, grupo, popularidade, nomeTecnologiaDestino, peso, conforme ilustrado no` **`exemplo de execução.`**

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:

```
./programaTrab
3 tecnologias.bin 2
nomeTecnologiaOrigem "AZURE"
nomeTecnologiaOrigem "ENTITY-FRAMEWORK"
AZURE, 2, 14, .NET, 22
ENTITY-FRAMEWORK, 2, 13, .NET, 3
```

[4] Permita a recuperação dos dados de um registro, a partir da identificação do RRN (número relativo do registro) do registro desejado pelo usuário. Por exemplo, o usuário pode solicitar a recuperação dos dados do registro de RRN = 2 ou do registro de RRN = 4. Registros marcados como logicamente removidos não devem ser exibidos.

Sintaxe do comando para a funcionalidade [4]:

```
4 arquivoEntrada.bin  RRN
```

onde:

- `arquivoEntrada.bin` é o arquivo binário de entrada, o qual foi gerado conforme as especificações descritas neste trabalho prático.
- `RRN` é o número relativo do registro desejado pelo usuário.

Saída caso o programa seja executado com sucesso:

Será recuperado, no máximo, 1 registro. Para esse registro, os valores de seus campos devem ser mostrados em uma única linha, sendo separados por vírgula. Caso o campo seja nulo, deve ser exibido: NULO. Ordem de exibição dos campos: nomeTecnologiaOrigem, grupo, popularidade, nomeTecnologiaDestino, peso, conforme ilustrado no **exemplo de execução.**

Mensagem de saída caso não seja encontrado o registro ou o registro esteja removido:

```
Registro inexistente.
```

Mensagem de saída caso algum erro seja encontrado:

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

Exemplo de execução:

```
./programaTrab
4 tecnologias.bin 1
SQL-SERVER, 2, 65, .NET, 33
```

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.

[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 no 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.
- Um vídeo gravado pelos integrantes do grupo, o qual deve ter, no máximo, 5 minutos de gravação. O vídeo deve explicar o trabalho desenvolvido. Ou seja, o grupo deve apresentar: cada funcionalidade e uma breve descrição de como a funcionalidade foi implementada. Todos os integrantes do grupo devem participar do vídeo, sendo que o tempo de apresentação dos integrantes deve ser balanceado. Ou seja, o tempo de participação de cada integrante deve ser aproximadamente o mesmo. O uso da webcam é obrigatório.

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 arquivo zip. Adicionalmente, no arquivo Makefile é importante se ter um *tab* nos locais colocados acima, senão ele pode não funcionar.

Instruções de entrega.
O programa deve ser submetido via [run.codes]:

- página: https://runcodes.icmc.usp.br/
- Código de matrícula: **Z3BS**

O vídeo gravado deve ser submetido por meio da página da disciplina no e-disciplinas, no qual o grupo vai informar o nome de cada integrante, o número do grupo e um link que contém o vídeo gravado. Ao submeter o link, verifique se o mesmo pode ser acessado. Vídeos cujos links não puderem ser acessados receberão nota zero. Vídeos corrompidos ou que não puderem ser corretamente acessados receberão nota zero.

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.
- Vídeo. Integrantes que não participarem da apresentação receberão nota 0 no trabalho correspondente.

Casos de teste no [run.codes]. Juntamente com a especificação do trabalho, serão disponibilizados 70% dos casos de teste no [run.codes], para que os alunos possam avaliar o programa sendo desenvolvido. Os 30% restantes dos casos de teste serão utilizados nas correções.

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 realização do trabalho prático com alunos de turmas diferentes 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).

Bom Trabalho!