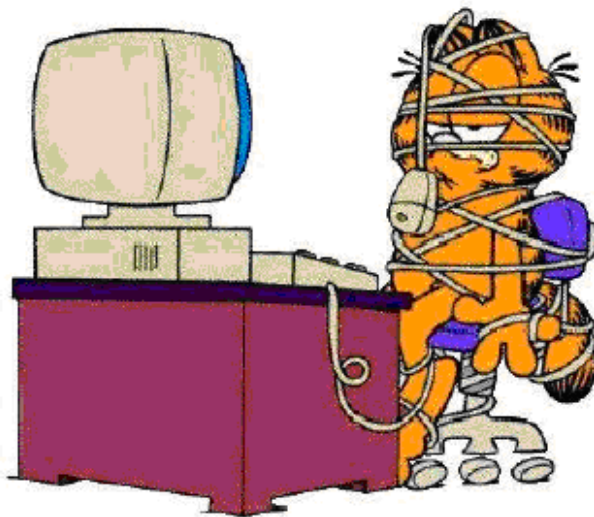
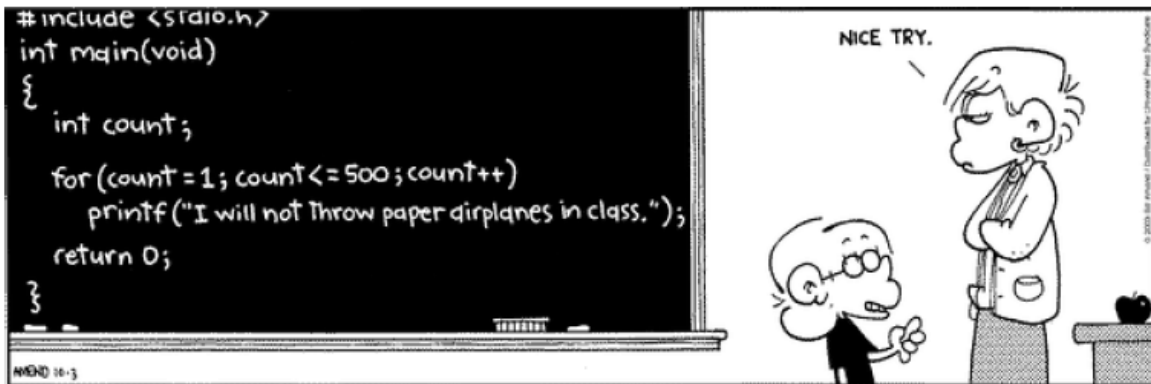


Certa vez, numa escola americana, uma menina levou o castigo de escrever 500 vezes no quadro que nunca mais iria jogar aviões de papel na sala de aula.
Ela é esperta.....



Vivendo e aprendendo...

Se mexer, pertence à Biologia.
Se feder, pertence à Química.
Se não funcionar, pertence à Física.
Se ninguém entende, é Matemática.
Se não faz sentido, é Economia ou Psicologia.
Se não mexe, não fede, não funciona,
ninguém entende e não faz sentido,
então é INFORMÁTICA...

CURIOSIDADE :

Para quem ainda tem dificuldade de saber a diferença entre Software e Hardware:

- Software: é a parte que você xinga.
- Hardware: é a parte que você chuta.

Capítulo

1

Conceitos Básicos e Terminologia



Introdução

O estudo da organização e do gerenciamento de arquivos é fundamental para uma maior eficiência na implantação de sistemas de arquivos, ou até mesmo em sistemas de BD.

É importante frisar que o tempo de acesso à um dado na memória externa é muito grande comparado com a da memória interna (milissegundos x microssegundos)

Portanto, faz-se necessário um estudo detalhado do tipo de organização devemos utilizar para cada tipo de aplicação, melhorando o acesso ao dado, dispondo para isto de definições de caminhos de acesso aos dados que tenham um desempenho satisfatório.

Conceitos Básicos

Alguns conceitos básicos são descritos a seguir.

Dado é qualquer parte de um certo conjunto que necessita de uma determinada gerência.

Por exemplo:

No mundo real, uma Universidade X, representa uma determinada realidade que hipoteticamente deve-se fazer uma gerência de determinados “objetos” que fazem parte dessa realidade.

A Figura 1.1 ilustra a ideia básica na captura dos dados de um determinado modelo sistêmico.

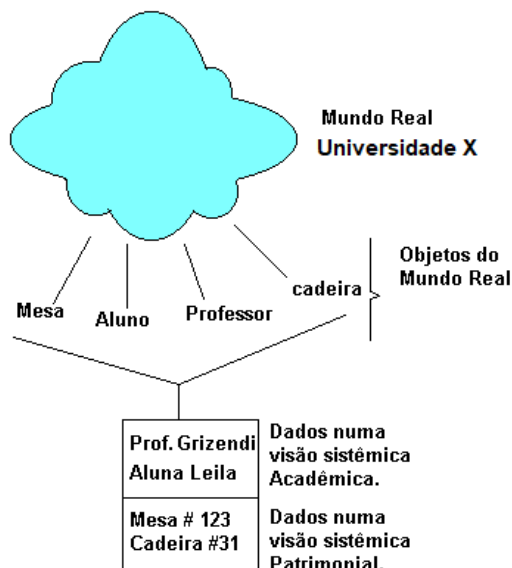


Figura 1.1 – Visão esquemática do mundo real e alguns dados.

Numa visão sistêmica acadêmica, o professor Grizendi, o professor Carlos Alberto, são elementos que desejamos gerenciar e armazená-los, logo, esses elementos serão considerados dados. Poderíamos armazenar todos os professores em um repositório específico(arquivo) da Universidade X. Os atributos dos professores a serem armazenados (nome, matrícula,...) são os dados de cada professor que serão armazenados.

A disciplina de ED visa estudar as mais diversas formas diferentes de se armazenar esses dados em um arquivo.

Os dados armazenados por características comum geram os arquivos, e os arquivos dentro de uma mesma visão sistêmica geram os Banco de Dados.

Quando os dados são manipulados por sistemas de informação dizemos que esses dados viram informações.

Memória

A **memória** é um conjunto de **posições de memória** endereçáveis. Uma **posição de memória** fisicamente é um **byte**, logo pode-se concluir que fisicamente que **memória** é um conjunto de **bytes**. Logicamente, a memória é um conjunto de caracteres, que agrupados formam um **dado** que pode ser uma palavra, um texto, um gráfico, um desenho, um áudio, vídeo etc.

Memória Principal / Secundária

A **memória** pode ser dividida em principal e secundária.

A **memória principal** é aquela em que o dado que nela estiver é processado diretamente pela **CPU**. Exemplos de **memória principal (MP)** temos CACHE, ROM, RAM, PROM, EPROM etc. Geralmente estas memórias são voláteis, isto é, quando o computador é desligado, os dados nela contida são perdidos.

A **memória secundária (MS)** é aquela em que o dado que nela estiver não é processado diretamente pela **CPU**, ele é transportado para a **memória principal**, via **canal**, para posterior processamento. Exemplos de **memória secundária** temos disco rígido (HD), discos flexíveis (disquetes), o CD, Zip Driver Disk, fitas magnéticas, fitas DAT ...???. Estas memórias são permanentes, isto é, os dados nela armazenados não são perdidos quando desligamos o computador.

Meios Físicos de Armazenamento

Diversos tipos de dispositivos de armazenamento de dados existem na maioria dos computadores. Esses meios de armazenamento são classificados pela velocidade de acesso aos dados, pelo custo por unidade de dados do dispositivo e pela confiabilidade.

Os dispositivos de armazenamento existem tanto para memória principal e memória secundária. A seguir descrevem-se alguns meios físicos de armazenamentos para **MP** e **MS**.

Memória principal

Este é o meio de armazenamento usado para as dados que estão disponíveis para ser operados diretamente pela CPU. As instruções de máquina de uso geral operam na memória principal.

Embora a memória principal possa conter diversos **megabytes** de dados, ela é geralmente muito pequena para armazenar o banco de dados inteiro. Alguns exemplos de **MP**: **RAM**, **ROM**, **PROM**, **EPROM**, **buffers**, **Cache** etc.

Memória secundária

Este é o meio de armazenamento usado para as dados que não estão disponíveis para ser operados diretamente pela CPU. Eles devem ser transferidos via canal e disponibilizados nos buffers de leitura/gravação.

Alguns exemplos de MS: Disco magnético, SSD, fita....

Meios Físicos de Armazenamento para MS

Disco Magnético

Este é o meio principal utilizado para o armazenamento de dados de longa permanência. Tipicamente, todo o banco de dados é armazenado em disco. Os **dados** precisam ser movidos do disco para **a memória principal** para que sejam processados. Após as operações, os dados precisam retornar ao disco. O armazenamento em disco é conhecido como armazenamento de acesso direto (**DASD – Device Access Storage Direct**) porque é possível ler os dados do disco em qualquer ordem (exceto nos discos de armazenamento de acesso sequencial como por exemplo as fitas magnéticas).

O armazenamento em disco normalmente sobrevive a falhas de energia elétrica e quedas do sistema (permanentes) . Os discos de armazenamento podem eles próprios falhar e destruir dados, mas tais falhas são significativamente menos frequentes do que as partes de sistema. A Figura 1.2 ilustra os elementos de um disco magnético.

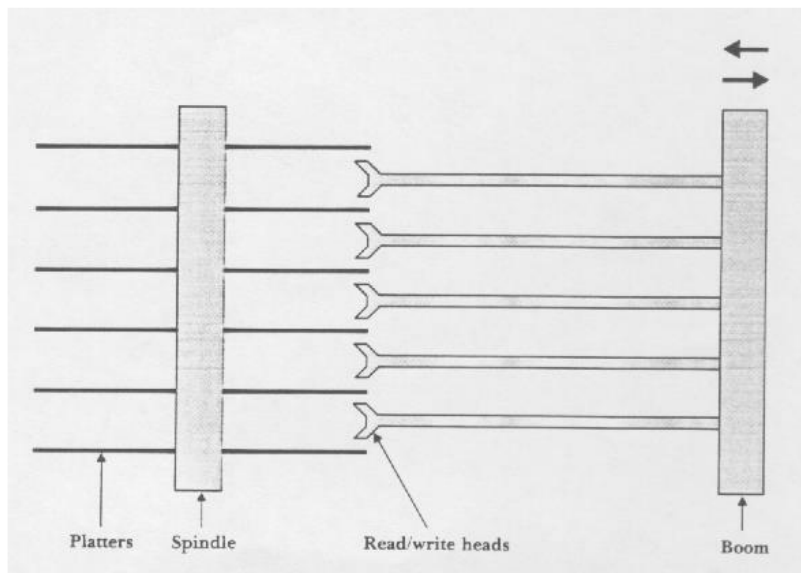


Figura 1.2 – Esboço de um corte em um disco magnético.

Cada *prato* (disco) é dividido em *trilhas* e cada *trilha* é dividida em *setores* e finalmente *cada setor* é dividido em *blocos* de registros. Os setores são separados por **gaps**. A figura 1.3 ilustra a superfície de um disco mostrando as trilhas e os setores.

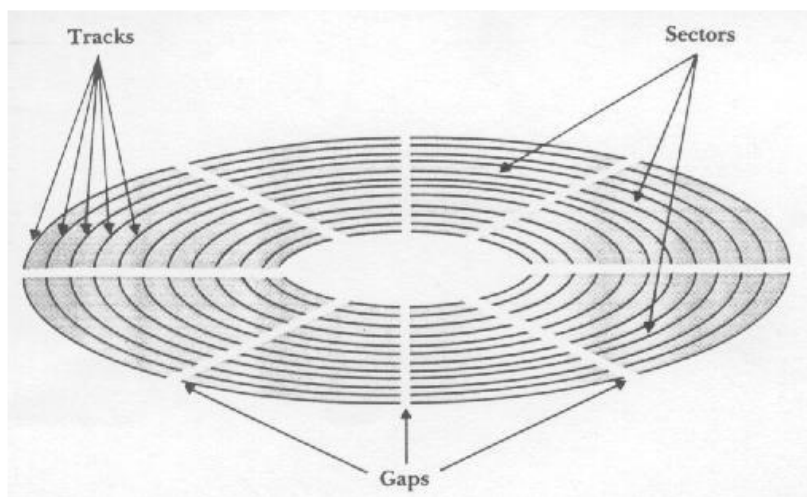


Figura 1.3 – Elementos da Superfície de um disco.

Os trilhas de mesmo raio definem logicamente os denominados **cilindros**. A figura 1.4 ilustra esquematicamente um dispositivo em disco com um conjunto de sete cilindros.

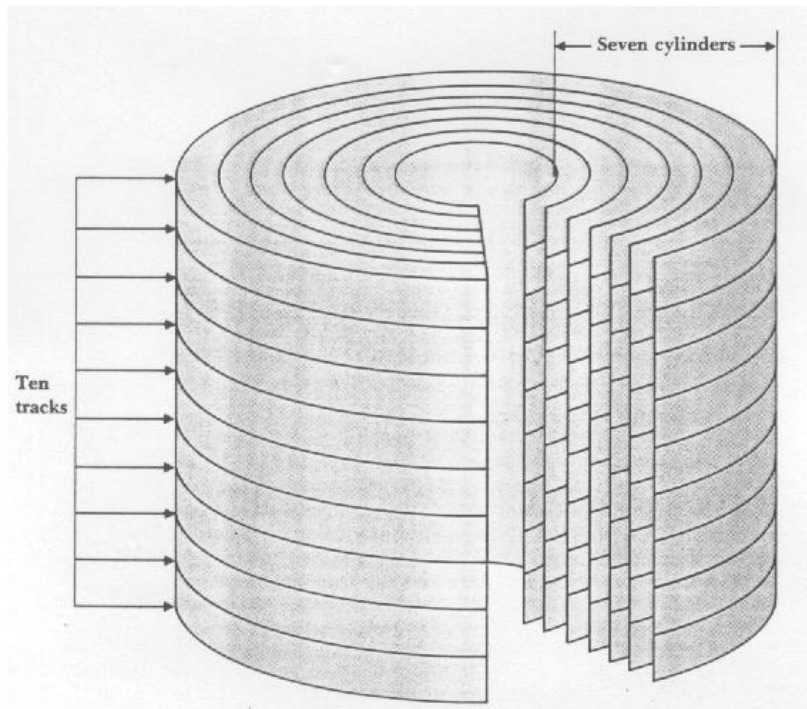
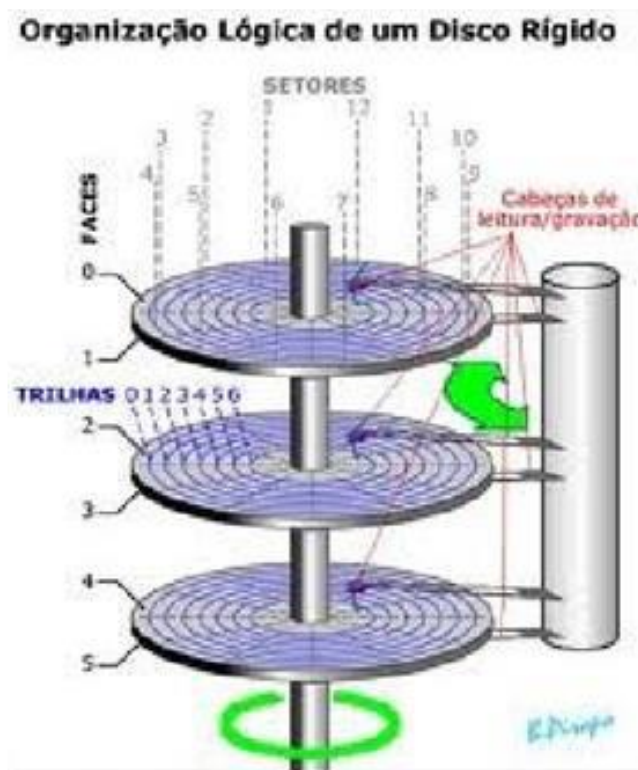


Figura 1.4 – Cilindro e Trilhas de um disco.



Fita Magnética

Este é o dispositivo de armazenamento usado primariamente para cópias de reserva e dados históricos. Embora a fita seja bem mais barata do que o disco, o acesso aos dados é muito lento, uma vez que a fita precisa ser lida sequencialmente a partir do início. Por esta razão, o armazenamento em fita é chamado de armazenamento de acesso sequencial e é usado primariamente para recuperar falhas do disco.

Os gravadores de fita são menos complexos do que os discos, portanto, mais confiáveis. A Figura 1.5 ilustra uma fita magnética, com os seus elementos básicos.

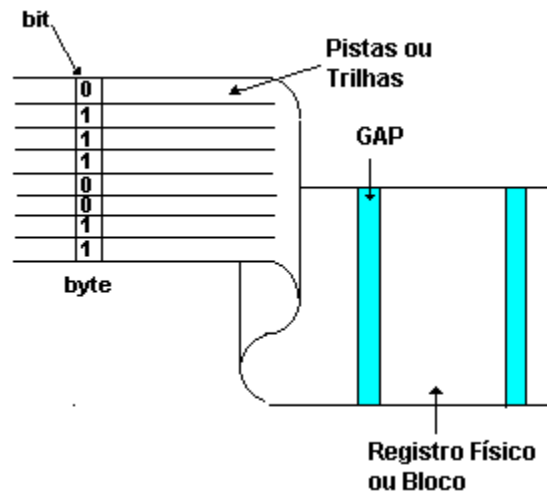


Figura 1.5 – Elementos em uma fita magnética.

A fita é dividida em conjunto de bytes constituindo um *bloco* ou *registro físico*. Dois *blocos* são separados por um **GAP**. Este separador contém informações como por exemplo o tamanho do bloco, número de registro lógicos existentes no bloco etc.

Um *bloco* ou um **registro físico** na realidade é um conjunto de **registros lógicos**, onde um registro lógico é um conjunto de informações lógicas armazenadas.

A fita é dividida em **trilhas** ou **pistas** (7 a 9) . A sequência de **bits** numa mesma vertical determina o **byte**. Cada **byte** tem como primeiro **bit** (ou último) um **bit** especial denominado de **bit de paridade**. Este **bit** serve para verificar se a leitura do **byte** pela unidade de fita foi correta.

A quantidade de **bytes** por polegada em uma mesma fita é denominada de **densidade**. A **densidade** é medida em **bpi** (bytes per inch). Por exemplo a densidade 6500 **bpi** significa que em uma polegada cabem 6500 *bytes*.

O número de **registros lógicos** em um mesmo **bloco** (**registro físico**) é denominado de **fator de bloco**. Isto é, se o **fator de bloco** de uma determinada fita é 20, isto significa que em cada **bloco** cabem 20 **registro lógicos**.

Estruturas Lógicas e Físicas

É muito importante sabermos diferenciar as estruturas lógicas das estruturas físicas. Uma Estrutura lógica é uma estrutura cuja visão é do programador enquanto a estrutura física é a visão do Sistema Operacional.

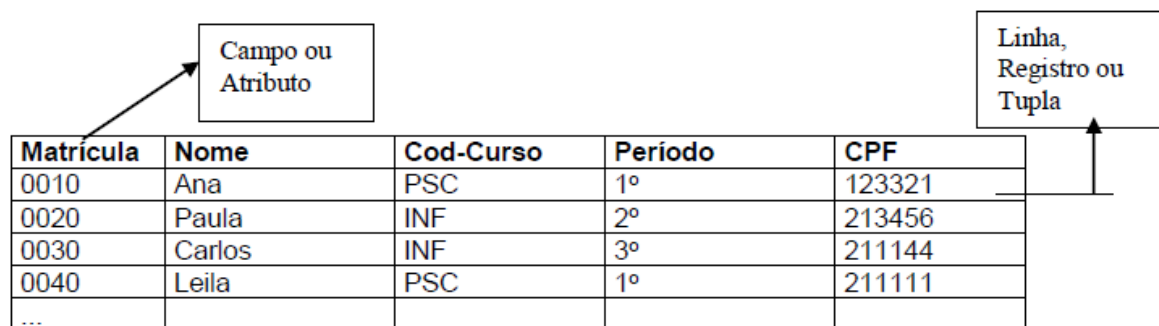
Por exemplo um caractere qualquer "A" é para um programador uma letra que representa uma parte de uma outra estrutura lógica qualquer. Este caractere representado fisicamente é representado utilizando um conjunto de **bits**, formando um **byte**, que é a estrutura física correspondente. Fisicamente, o caractere A está representado na forma binária em ASC II como o conjunto de 8 bits 10001001.

A seguir descrevem-se as principais estruturas lógicas e suas correspondências físicas.

Estruturas Lógicas

As estruturas lógicas são aquelas cujo domínio de visão é a do programador. Por exemplo, seja uma *tabela* contendo as informações acadêmicas sobre os alunos da Faculdade X ilustrada na Tabela 1 a seguir. Esta *tabela*, definida logicamente, é denominada de *Arquivo Lógico*. As linhas da tabela são denominadas de *registros lógicos* ou *tuplas*. As colunas são denominadas de *campos* ou *atributos*.

Tabela 1.1 – Exemplo da estrutura lógica Arquivo.



The diagram shows a table with five columns: Matricula, Nome, Cod-Curso, Período, and CPF. The first four columns have a label 'Campo ou Atributo' with an arrow pointing to the 'Nome' column. The last column has a label 'Linha, Registro ou Tupla' with an arrow pointing to the last row of the table.

Matricula	Nome	Cod-Curso	Período	CPF
0010	Ana	PSC	1º	123321
0020	Paula	INF	2º	213456
0030	Carlos	INF	3º	211144
0040	Leila	PSC	1º	211111
...				

Pode-se concluir que uma *tabela* é uma coleção de *linhas*. Cada *linha* é um conjunto de *campos* e finalmente, cada *campo* é um conjunto de *caracteres*. A figura 1.6 ilustra a hierarquia de conceitos lógicos.

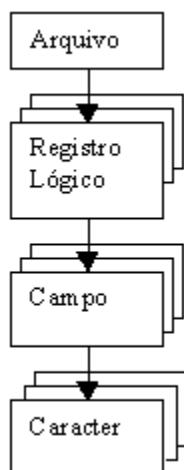


Figura 1.6 – Hierarquia de conceitos Lógicos.

Registro Lógico

Um arquivo contém uma coleção de um ou mais registros lógicos. Antes de definirmos o arquivo precisamos criar a estrutura do registro deste arquivo, isto é, quais campos formam cada registro. A definição em C é utilizando o comando **struct**.

Sintaxe : **struct** < nome do registro >
{
 tipo da variável < variável 1 > ;
 tipo da variável < variável 2 > ;
 ...
 tipo da variável < variável n > ;
};

Exemplo :

Considere um registro com o seguinte leiaute:

Matrícula	Nome	Cod-Curso
-----------	------	-----------

Definição da Estrutura de Dados em C:

```
struct REG_ALUNO  
{  
    int MATRICULA;  
    char NOME[30];  
    char CODCURSO[3];  
};
```

Arquivo Lógico

As **tabelas** ou **arquivos lógicos** podem ser gravados em memória principal ou secundária (fita ou discos). Para estudarmos os métodos de ordenação interna, utilizamos as tabelas gravadas em memória principal que nada mais nada menos é um vetor de registros. Consequentemente, no estudo de métodos de ordenação externa, utilizamos as tabelas gravadas em memória secundária.

Considere a tabela 2 a seguir representa um arquivo lógico denominado por exemplo de **Aluno**.

Tabela 1.2 – Exemplo de uma Tabela de Alunos com 3 campos.

Aluno			
No Registro	Matrícula	Nome	Cod-Curso
0	0010	Ana	PSC
1	0020	Paula	INF
2	0030	Carlos	INF
3	0040	Leila	PSC
...	...		

Arquivo em Memória Secundária

Supondo que este arquivo esteja num diretório "c:\arquivos", considerando o sistema operacional MS Windows, o **arquivo lógico** corresponderá ao **arquivo físico** " c:\arquivos\alunos.dat".

Chave

Para classificarmos um determinado arquivo e melhorar o acesso aos seus registros precisamos definir alguns conceitos relativos aos campos que fazem parte da classificação.

- Chave Primária

É um conjunto de um ou mais campos que definem univocamente uma e somente uma linha de uma tabela.

Na tabela 1 os campos Matrícula e CPF poderiam ser chaves primárias, pois, cada linha desta tabela tem matrículas diferentes e CPFs também diferentes.

A chave primária deve ser "dataless", isto é, sem significado. Mudança de significado pode implicar na mudança do valor da chave. "Ana", "ANA", ou "ana" devem levar ao mesmo registro.

Formas canônicas para as chaves: representação obedece uma regra

- Chave Secundária

É um conjunto de um ou mais campos que definem uma ou mais linhas de uma tabela.

Na tabela 1 o campo Nome é um bom exemplo de chave secundária.

- Chave Candidata

É um conjunto de um ou mais campos que definem univocamente uma e somente uma linha de uma tabela, mas não foi escolhida como chave primária por uma visão sistêmica diferente.

Na tabela 1 os campos Matrícula e CPF poderiam ser chaves candidatas, pois, cada linha desta tabela tem matrículas diferentes e CPFs também diferentes.

Numa visão sistêmica Acadêmica a chave primária seria Matrícula e a candidata o CPF, e numa visão sistêmica Financeira, a chave primária seria o CPF e a candidata o Matrícula.

- Chave Estrangeira

É um conjunto de um ou mais campos que em outra tabela são chaves primárias.

Na tabela 1 o campo Cod-Curso é exemplo de chave estrangeira, pois existe uma outra tabela de cursos que se relaciona com a tabela de alunos pela chave Cod-Curso. Pode-se observar que o campo Cod-Curso na outra tabela 3 é uma chave primária.

Tabela 1.3 – Exemplo de uma Tabela de Cursos.

Cod-Curso	Descrição	No Créditos	No Períodos	...
PSC	Psicologia
INF	Informática			
LET	Letras			
ENG	Engenharia			
...	...			

A figura 1.7 ilustra o relacionamento entre as tabelas Aluno (Tabela 1.1) e Curso (Tabela 1.2).

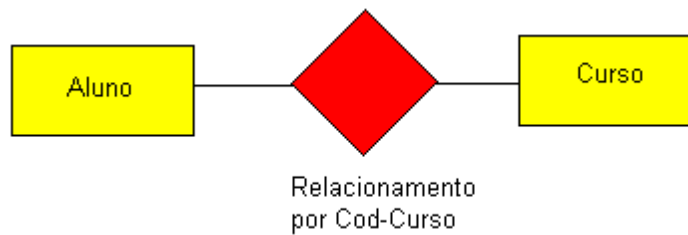


Figura 1.7 – Hierarquia de conceitos Lógicos.

EMPREGADO

Chave de pesquisa: 1030

Matricula	Nome	Idade	Salário
1000	Ademar	32	5000
1010	Roberto	25	7500
1020	Gerson	43	6000
1030	Yeda	23	9000
1040	Bernardo	21	4500
1050	Ângela	29	5000

Chave de ordenação

Estruturas Físicas

As estruturas físicas são aquelas cujo domínio de visão é a do SO. A seguir descrevem-se alguns conceitos físicos importantes.

Bit.

É a menor determinação física.

Byte.

É um conjunto de bits.

Setor

É um conjunto de blocos numa mesma superfície de um disco.

Trilha.

É um conjunto de setores que possuem mesmo raio numa mesma superfície de um disco.

Cilindro

É um conjunto de trilhas de mesmo raio em superfícies de discos diferentes superpostos.

Bloco ou Registro Físico

É um conjunto de bytes que pode ser exemplificada em um setor, uma trilha ou até mesmo um cilindro.

Data Set

É um conjunto ou uma coleção de registros físicos, contendo no mínimo de um cilindro.

A figura 1.8 ilustra a hierarquia de conceitos físicos.

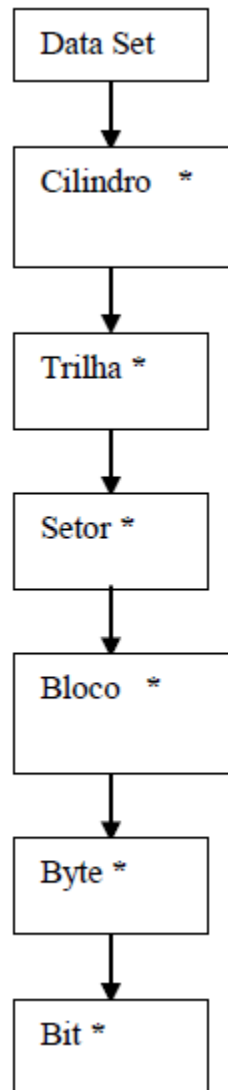
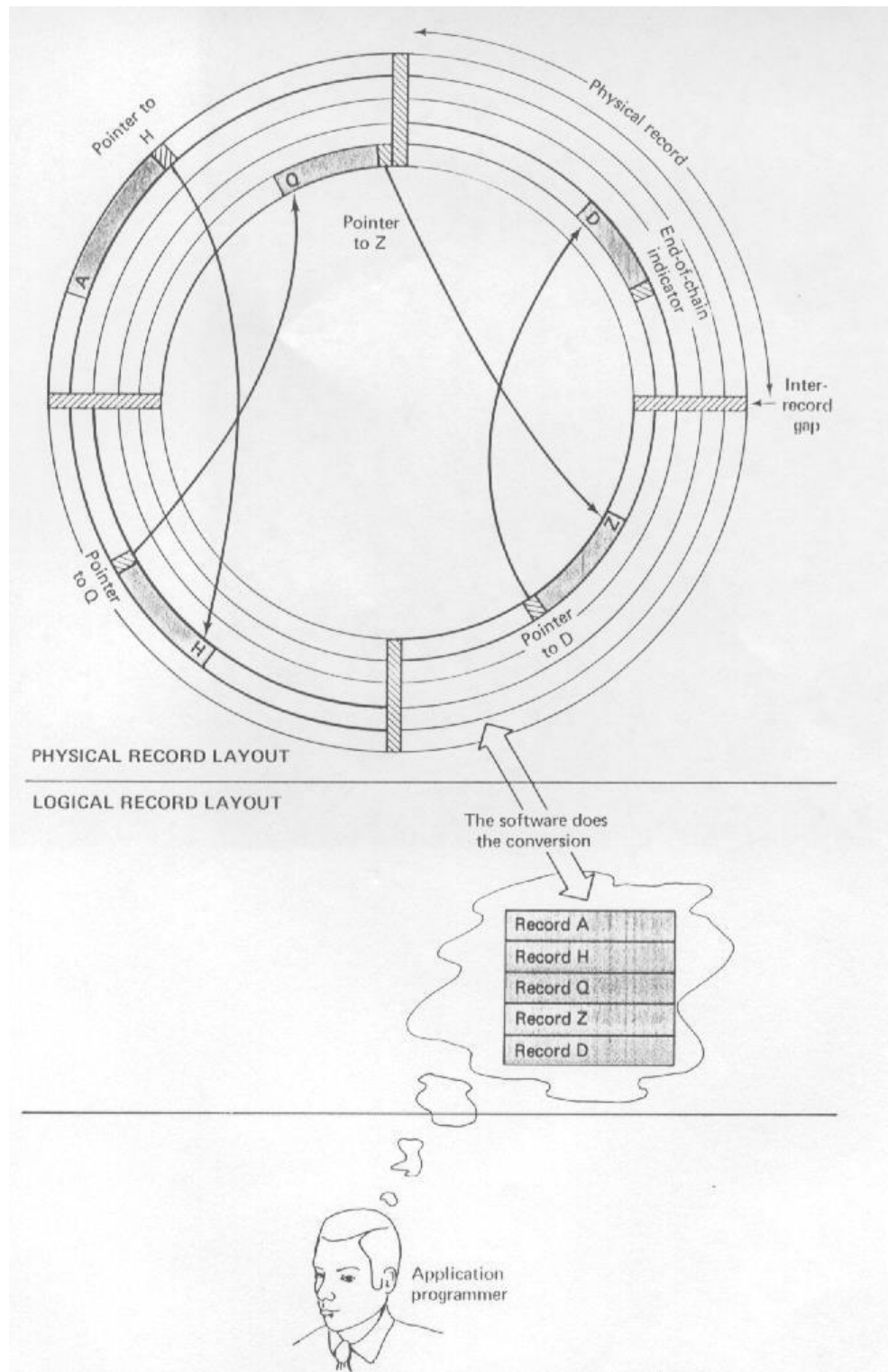
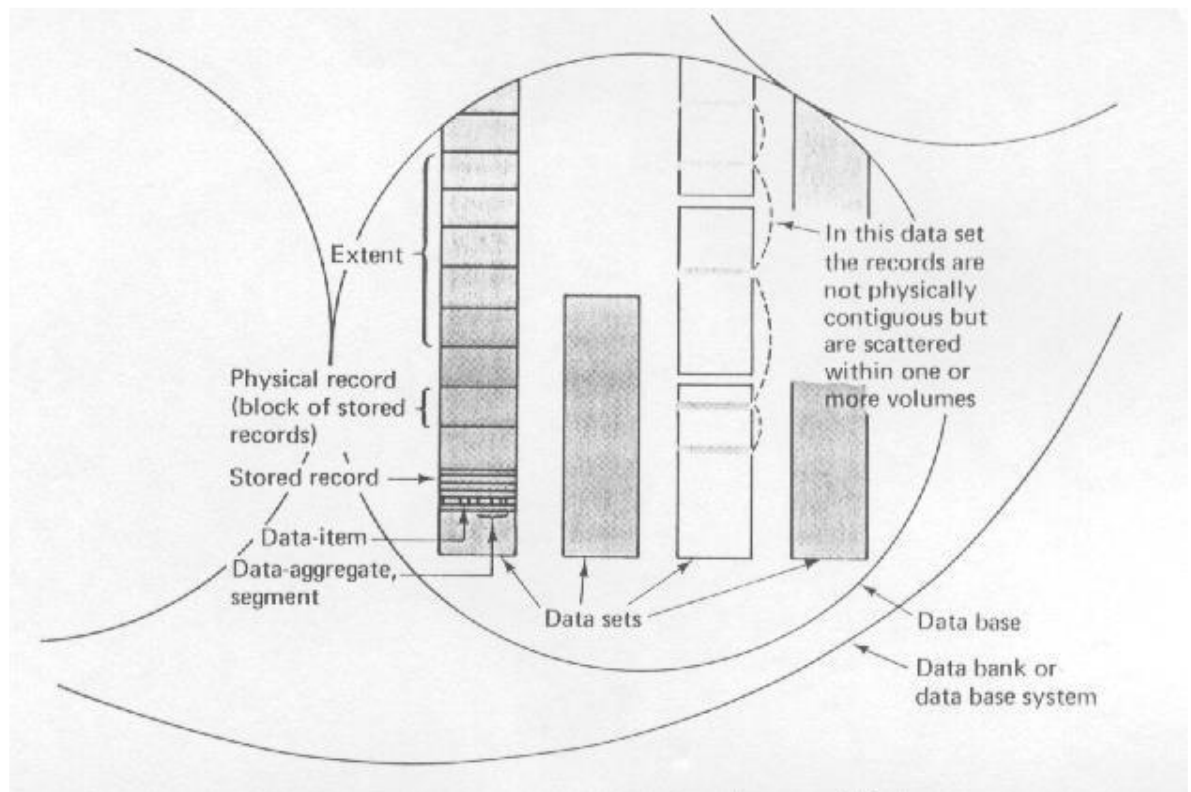


Figura 1.8 – Hierarquia de conceitos Lógicos.

A figura 1.9 ilustra a correspondência entre estruturas lógicas e estruturas físicas.



A figura 1.10 ilustra o armazenamento físico de dados. Pode-se observar os elementos físicos ilustrados hipoteticamente e os elementos lógicos correspondentes.



Tipos de armazenamento de Campos

➤ Comprimento fixo

- Cada campo tem tamanho predeterminado
- Recuperação facilitada
- Espaço alocado e não usado = **desperdício**
- Ruim para campos de dados com tamanho variável
- Razoável quando comprimento fixo ou com pouca variação

Maria	Rua 1	123	São Carlos
João	Rua A	255	Rio Claro
Pedro	Rua 10	56	Rib. Preto

➤ Indicador de comprimento

- Tamanho de cada campo antes do dado
- Se tamanho do campo < 256 bytes
- um único byte para indicar o comprimento

05Maria05Rua 10312310São Carlos
04João05Rua A0325509Rio Claro
05Pedro06Rua 10025610Rib. Preto

➤ Delimitadores

- Caracteres especiais inseridos ao final de cada campo
- Ex.: /, tab, #, etc...
- Espaços em branco geralmente não servem

Maria|Rua 1|123|São Carlos|
João|Rua A|255|Rio Claro|
Pedro|Rua 10|56|Rib. Preto|

➤ Uso de tags

- Vantagem: informação (semântica)
- Facilidade de identificação de conteúdo do arquivo
- Facilidade de identificação de campos perdidos
- Possibilidade de padronização (html, XML, ...)
- **Desvantagem:** keywords ocupam espaço

Nome=Maria|Endereço=Rua 1|Número=123|Cidade=São Carlos|
Nome=João|Endereço=Rua A|Número=255|Cidade=Rio Claro|
Nome=Pedro|Endereço=Rua 10|Número=56|Cidade=Rib. Preto|

Registros de tamanho fixo

- Todos os registros têm o mesmo número de bytes
- Muito comum
- Possível registros de tamanho fixo com campos de tamanho variável

Registro de tamanho fixo e campos de tamanho fixo:

Maria	Rua 1	123	São Carlos
João	Rua A	255	Rio Claro
Pedro	Rua 10	56	Rib. Preto

Registro de tamanho fixo e campos de tamanho variável:

Maria		Rua 1		123		São Carlos		← Espaço vazio →
João		Rua A		255		Rio Claro		← Espaço vazio →
Pedro		Rua 10		56		Rib. Preto		← Espaço vazio →

- Ao invés de números fixo de bytes, número fixo de campos
- O tamanho do registro é variável
- Campos separados por delimitadores

Registro com número fixo de campos:

Maria|Rua 1|123|São Carlos|João|Rua A|255|Rio Claro|Pedro|Rua 10|56|Rib. Preto|

Registros de tamanho variável

Indicador de tamanho para registros

- O indicador que precede o registro fornece o seu tamanho total, em bytes
- Os campos são separados internamente por delimitadores
- Boa solução para registros de tamanho variável

Registro iniciados por indicador de tamanho:

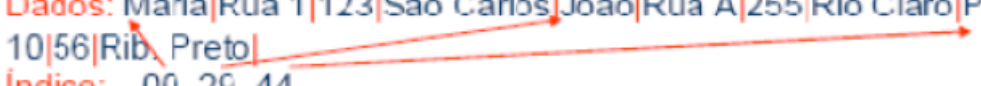
28Maria|Rua 1|123|São Carlos|25João|Rua A|255|Rio Claro|27Pedro|Rua
10|56|Rib. Preto|

Utilizar um índice

- Um índice externo poderia indicar o deslocamento de cada registro relativo ao início do arquivo
- Pode ser utilizado também para calcular o tamanho dos registros
- Os campos seriam separados por delimitadores...

Arquivos de dados + arquivo de índices:

Dados: Maria|Rua 1|123|São Carlos|João|Rua A|255|Rio Claro|Pedro|Rua
10|56|Rib. Preto|
Índice: 00 29 44

A diagram with three red arrows pointing from the index values to the start of each record in the data file. The first arrow points from '00' to the start of 'Maria|Rua 1|123|São Carlos|'. The second arrow points from '29' to the start of 'João|Rua A|255|Rio Claro|'. The third arrow points from '44' to the start of 'Pedro|Rua 10|56|Rib. Preto|'.

Utilizar delimitadores

- Separar os registros com delimitadores análogos aos de fim de campo
- O delimitador de campos é mantido, sendo que o método combina os dois delimitadores
- Note que delimitar fim de campo é diferente de delimitar fim de registro

Registro delimitado por marcador (#):

Maria|Rua 1|123|São Carlos|#João|Rua A|255|Rio Claro|#Pedro|Rua 10|56|Rib.
Preto|