Data Engineering

Filipe Nascimento
fgvn@cesar.school

C.E.S.A.R

C . E . S . A . R

Pessoas impulsionando inovação.

Inovação impulsionando negócios.

Everton Dias
etgdb@cesar.org.br

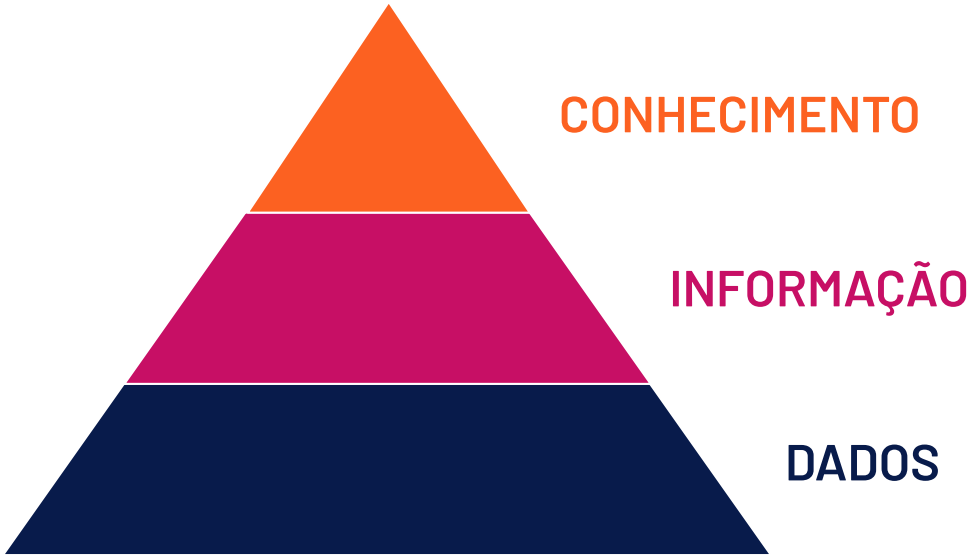
Janaína Branco
jcb@cesar.org.br

BANCOS DE DADOS

Os bancos de dados são parte fundamental do cotidiano das empresas e seus clientes, sendo um dos principais objetos de trabalho do engenheiro de dados, hoje iremos dar uma introdução aos bancos de dados.

BANCOS DE DADOS

DADOS, INFORMAÇÃO E CONHECIMENTO



DADOS, INFORMAÇÃO E CONHECIMENTO

O DADO representa a menor porção da informação, não possuindo significado relevante em si mesmo tampouco possuindo a capacidade de conduzir a algum sentido, não podendo apoiar nenhuma decisão sozinho muito mesmo confirmar ou afirmar algo.

Mesmo possuindo vários dados, se estes não possuírem nenhum tipo de organização serão inúteis pois será impossível obter alguma INFORMAÇÃO a partir destes.

DADOS, INFORMAÇÃO E CONHECIMENTO

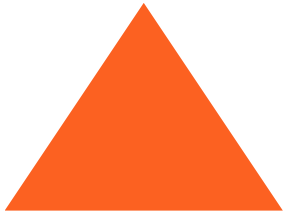
INFORMAÇÃO

A INFORMAÇÃO é o produto obtido a partir da combinação organizada dos dados de maneira a transmitir significado e compreensão dentro de um determinado contexto.

Quando a partir de de um conjunto de DADOS podemos extrair alguma INFORMAÇÃO, conseguimos gerar o CONHECIMENTO.

DADOS, INFORMAÇÃO E CONHECIMENTO

CONHECIMENTO



A partir das INFORMAÇÕES é possível construir o CONHECIMENTO que será a matéria prima das tomadas de decisões entre outras aplicações que culminam no atendimento das necessidades corporativas, individuais ou coletivas.

DADOS, INFORMAÇÃO E CONHECIMENTO

Exemplo

BATEU

VERMELHO

GANHOU

META

FERRARI

JOÃO

DADOS, INFORMAÇÃO E CONHECIMENTO

Exemplo

JOÃO BATEU A META E GANHOU UMA FERRARI VERMELHA

DADOS, INFORMAÇÃO E CONHECIMENTO

Exemplo

MANGA

DADOS, INFORMAÇÃO E CONHECIMENTO

Exemplo

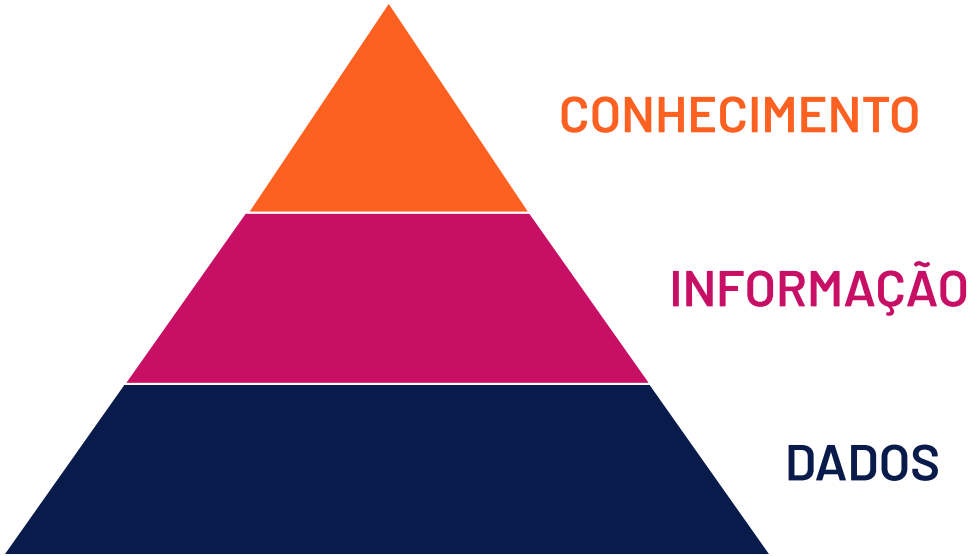
MANGA ***ROSA R$ 10,00 O KG***

MANGA

CONTEXTO

RICARDO DOBROU A **MANGA** ***DA CAMISA***

DADOS, INFORMAÇÃO E CONHECIMENTO



DADOS, INFORMAÇÃO E CONHECIMENTO

DADOS

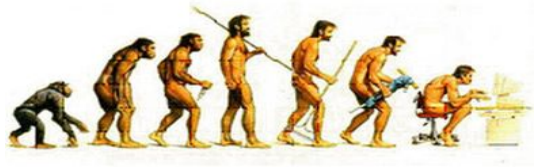
INFORMAÇÃO

CONHECIMENTO

BANCOS DE DADOS

Antigamente as empresas costumavam armazenar seus dados em fichas de papel que por sua vez eram organizadas em pastas, sendo extremamente difícil e custoso a extração e manutenção destes arquivos. Em um passo seguinte estes arquivos evoluíram de físicos para digitais.

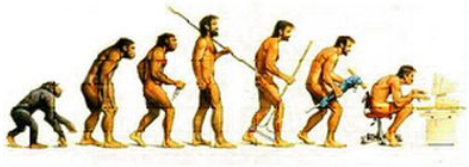
Inicialmente, cada entidade ficava em um arquivo de dados que acompanhado de um software permitia sua manipulação, trazendo uma melhoria dado o cenário anterior porém não era o suficiente e as necessidades logo trouxeram a tona o fato de que as entidades precisavam se relacionar.



BANCOS DE DADOS

Na década de 60 a gigante IBM investiu pesadamente em pesquisas para solucionar estes problemas dos bancos de dados digitais primitivos, surgindo nessa época vários modelos de bancos como o hierárquico entre outros. No final da década de 60 surge o primeiro SGBD - Sistema Gerenciador de Banco de Dados comercial.

Em 1970, Edgar Frank Codd da IBM



BANCOS DE DADOS

Com a evolução dos SGBDs saindo de sistemas de arquivos de armazenamento em disco, passando pela criação de novas estruturas mirando o armazenamento cada vez mais eficiente de informações e com o tempo passando a utilizar diferentes modelos de dados, servindo para descrever a estrutura das informações contidas nos seus bancos de dados.

- MODELO HIERÁRQUICO
- MODELO EM REDE
- MODELO RELACIONAL
- MODELO ORIENTADO A OBJETOS
- OBJETO-RELACIONAIS

BANCOS DE DADOS

NoSQL - Não aprofundaremos no momento mas vale a pena comentar sobre a continuidade destas importantes ferramentas de gerenciamento de bancos de dados que culminou em outros modelos de representação das estruturas de bancos de dados.

- MODELO DOCUMENTO
- MODELO CHAVE-VALOR
- MODELO GRAFO
- MODELO COLUNAR

BANCOS DE DADOS

MODELO RELACIONAL

Tendo como base os estudos teóricos realizados por Codd, o modelo relacional surgiu como um modelo mais flexível entregando maior independência entre os dados nos SGBDs provendo um conjunto de funções baseadas em álgebra relacional para recuperação e armazenamento de dados, permitindo processamento ***ad hoc***.
Em seu cerne fundamental está a estrutura conhecida como **RELAÇÃO** (TABELA)
que por sua vez é constituída de um ou mais **ATRIBUTOS** (CAMPOS) onde cada coluna possui um tipo de dados a ser armazenado, cada linha dessa tabela ou esquema é chamada de **TUPLA** (REGISTRO).

Devido a sua flexibilidade o modelo de dados não possui um caminho pré-definido para acessar os dados como seus antecessores.

BANCOS DE DADOS

MODELO RELACIONAL

O MODELO RELACIONAL irá representar as estruturas de dados organizadas em relações, porém para evitar problemas no uso deste paradigma é importante se atentar a algumas restrições como repetição de informações, perda de aspectos importantes para a construção da informação.

Algumas restrições são:

- Integridade Referencial
- Chaves
- Integridade de Junção de Relações

BANCOS DE DADOS

MODELO RELACIONAL

RELAÇÃO (TABELA) - ATRIBUTO(CAMPO) - TUPLA(LINHA)

codigo_usuario	nome_usuario	data_nascimento	indicador_ativo
5645683	João dos Santos	13/09/1986	True
0005452	Austregésilo Alves	02/02/1954	True
0000003	Ricardo das Dores	30/03/1935	False

BANCOS DE DADOS

MODELAGEM DE DADOS

A MODELAGEM DE DADOS é disciplina fundamental como parte da modelagem de sistemas, para o Engenheiro de Dados é indispensável o conhecimento sobre os conceitos da modelagem de dados seja ela lógica ou física.

Para isso entenderemos alguns conceitos como MODELAGEM CONCEITUAL, MODELAGEM LÓGICA e MODELAGEM FÍSICA.

BANCOS DE DADOS

MODELAGEM DE DADOS

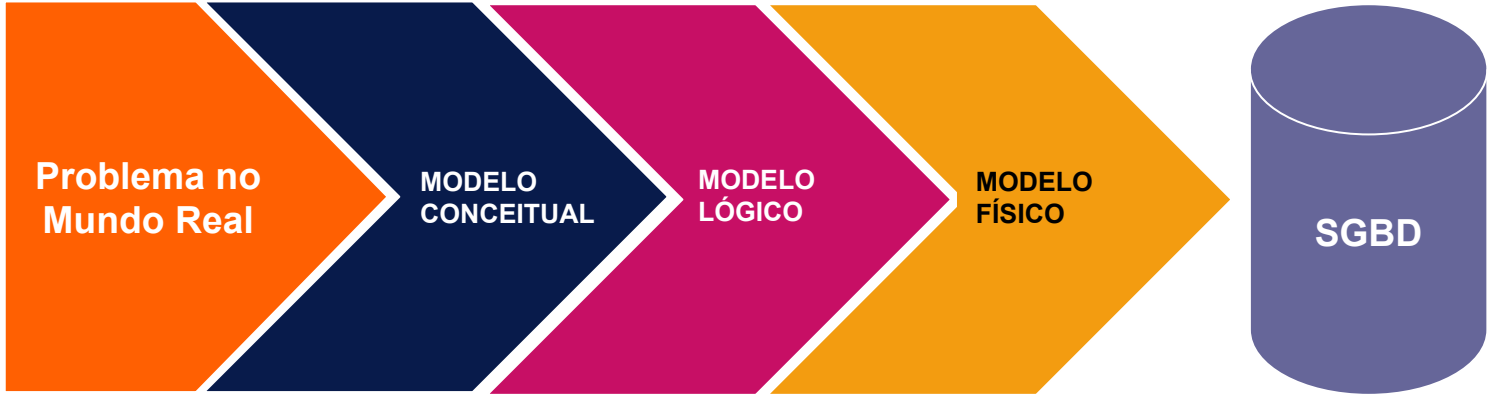
MODELAGEM CONCEITUAL - Tem como alvo a construção de um modelo conceitual de dados, definindo as ENTIDADES e seus RELACIONAMENTOS BÁSICOS;

MODELAGEM LÓGICA - Busca construir um modelo que demonstre as ligações entre as entidades e a lógica destas ligações;

MODELAGEM FÍSICA - Nesta forma o alvo já se torna a demonstração de como os dados deverão ser organizados fisicamente, como será construido e implementado no SGBD para garantir que o modelo lógico seja respeitado.

BANCOS DE DADOS

MODELAGEM DE DADOS



BANCOS DE DADOS

MODELAGEM DE DADOS

O modelo ENTIDADE-RELACIONAMENTO é um modelo de alto nível que não possui nenhuma dependência com sua escolha de SGBD, cujo objetivo é apenas a representação do problema a ser modelado. Sua representação pode ser feita através de um Diagrama de Entidade Relacionamento (DER), onde os retângulos representam Entidades, os losangos representam os relacionamentos entre as entidades.

As entidades são descritas a partir de seus atributos e que devem possuir um ou mais atributos que combinados representem a unicidade daquela instância (linha), a este atributo ou atributos chamados de chave primária ou primary key (PK) e que nunca podem ser nulos ou repetir-se em uma mesma entidade.

BANCOS DE DADOS

MODELAGEM DE DADOS

ENTIDADE - Objeto que pode ser distintamente identificado de outro objeto por um conjunto espefíco de atributos, podendo ser concreto ou abstrato.

BANCOS DE DADOS

MODELAGEM DE DADOS

RELACIONAMENTO - É a associação entre duas ou mais entidades.

BANCOS DE DADOS

MODELAGEM DE DADOS

CARDINALIDADE ou GRAU DE RELACIONAMENTO- Define o número máximo ou mínimo de ocorrências em uma entidade em relação a outra dentro de um relacionamento, onde as cardinalidades máximas são definidas pela letra n ou pelo número 1 (muitos). Cardinalidade mínima, 0 ou 1 são as consideradas.

BANCOS DE DADOS

MODELAGEM DE DADOS

1:1 - Acontece quando para cada ocorrência de uma entidade (gerente), temos somente uma ocorrência na entidade relacionada (setor).

BANCOS DE DADOS

MODELAGEM DE DADOS

1:n - Acontece quando para cada ocorrência de uma entidade (professor), temos várias ocorrências na entidade relacionada (turma).

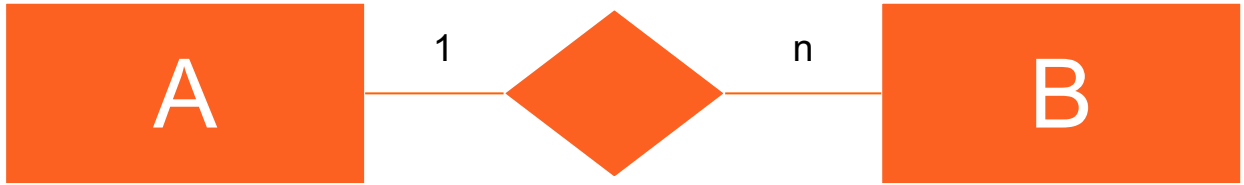
BANCOS DE DADOS

MODELAGEM DE DADOS

n:n - Ocorre quando não há restrições de ocorrências na formação das associações, ou seja, uma entidade pode estar associada a N ocorrências de outra entidade relacionada e vice e versa.

BANCOS DE DADOS

MODELAGEM DE DADOS



Entidade A está associada a zero (opcional) ou mais instâncias da Entidade B, que por sua vez está associada a uma (obrigatoriedade), e somente uma, instância da Entidade A. A esta notação chamamos de cardinalidade, onde o primeiro elemento indica a participação (opcional ou obrigatório) do relacionamento, enquanto o segundo representa o grau do relacionamento (um ou muitos).

BANCOS DE DADOS

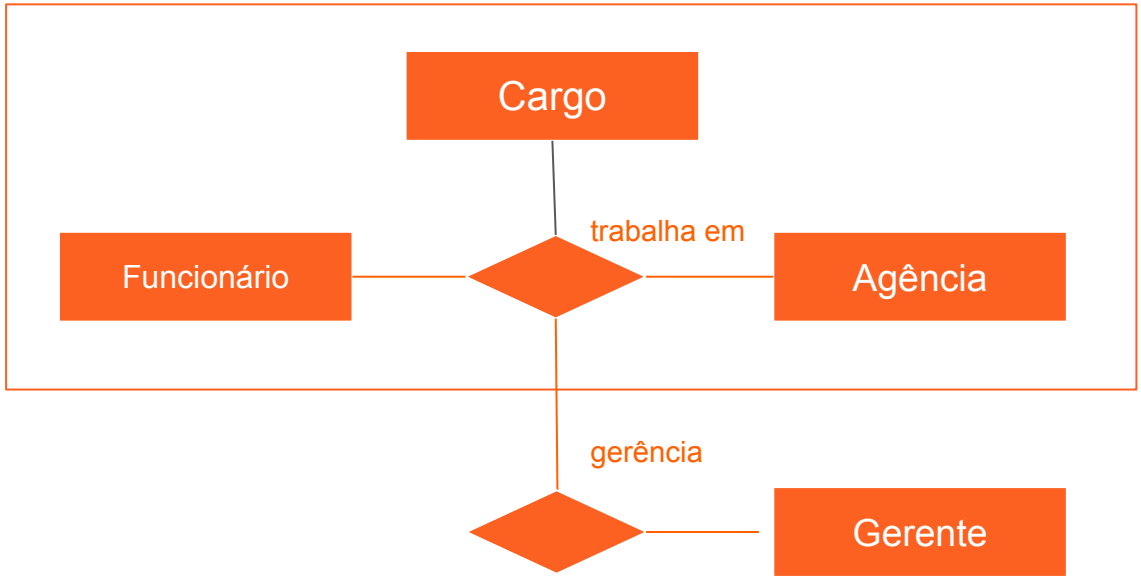
MODELAGEM DE DADOS

Existem ainda outros elementos para a construção do diagrama como a AGREGAÇÃO, RELACIONAMENTO TERNÁRIO, AUTO-RELACIONAMENTO e GENERALIZAÇÃO / ESPECIALIZAÇÃO.

BANCOS DE DADOS

MODELAGEM DE DADOS

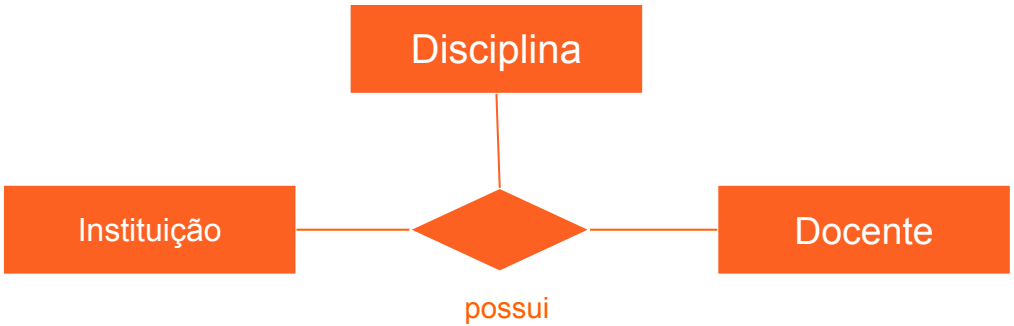
AGREGAÇÃO



BANCOS DE DADOS

MODELAGEM DE DADOS

RELACIONAMENTO TERNÁRIO



BANCOS DE DADOS

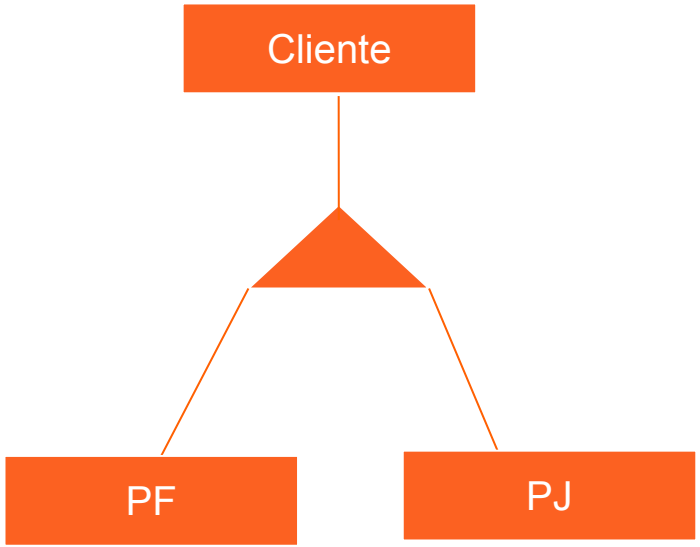
MODELAGEM DE DADOS

AUTO RELACIONAMENTO

BANCOS DE DADOS

MODELAGEM DE DADOS

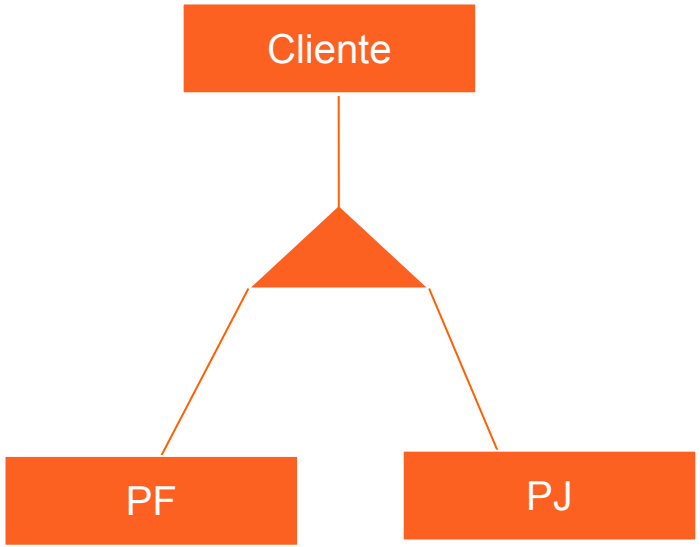
ESPECIALIZAÇÃO



BANCOS DE DADOS

MODELAGEM DE DADOS

ESPECIALIZAÇÃO



BANCOS DE DADOS

MODELAGEM DE DADOS

ATRIBUTOS

PRÁTICA

EXERCÍCIO 1

A FACULDADE

Uma faculdade contratou você como engenheiro de dados e pediu seu apoio para a construção de um modelo de banco de dados para gerenciar seus funcionários, divididos entre administrativos e professores, permitindo também alocar um professor por matéria por semestre, permitindo assim a rotação entre os professores durante os semestres.
O sistema também deve registrar os dados de matrícula dos alunos, o ano letivo e suas matérias que serão cursadas por semestre, vinculando aos professores que irão ministrar aulas nesse período.

Os dados referentes aos alunos devem conter sua filiação, sua data de nascimento, endereço completo, telefone para contato, e-mail e seu histórico de notas por matérias.

EXERCÍCIO 2

A ACADEMIA

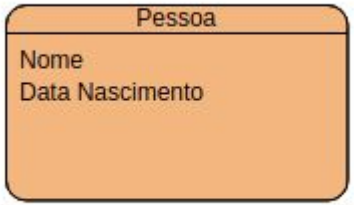
O dono da DATAFIT importante rede de academias lhe contratou e na primeira semana pediu seu apoio para revisar seu modelo de dados que realiza a gestão das academias, que deve possuir registro das unidades, seus funcionários e seus cargos, registro do horário de trabalho.

Deve também ter registros dos alunos e seus respectivos professores em caso de contratação de pacote individual que estará livre da regra de turnos, a regra de turnos diz que cada professor fará acompanhamento dos alunos sem exclusividade em uma unidade apenas por um turno por dia, podendo estar em outro turno em outra unidade.
O cadastro dos alunos deve permitir a identificação individual deles, endereço e formas de contato, também deve registrar o tipo do plano e a forma de pagamento.

BANCOS DE DADOS

MODELAGEM LÓGICA

Nesta etapa buscaremos transformar o modelo conceitual obtido na fase anterior em um modelo lógico, que definirá a implementação de como o banco de dados será construído.



BANCOS DE DADOS

MODELAGEM LÓGICA

TABELAS: Estrutura de Armazenamento das informações da Base de Dados.

REGISTROS: Os registros são a essência dos dados, ou seja a informação real que está armazenada na Tabela.

CAMPOS: São características que definem o dado a ser armazenado. O nome do Registro.

CHAVES: A forma de conexão entre as Tabelas. O relacionamento é definido por meio das Chaves.

PRIMARY KEY: Ou chave primária. É a principal Chave da Tabela. Toda tabela deve ter a sua.

SECONDARY KEY(s): São chaves secundárias que possibilitam a recuperação de informações (registros) das tabelas.

FOREIGN KEY: Uma chave estrangeira (FK) é uma coluna ou combinação de colunas que é usada para estabelecer e impor um link entre os dados em duas tabelas. Você pode criar uma chave estrangeira definindo uma restrição FOREIGN KEY ao criar ou modificar uma tabela.

ÍNDICES: Os dados são armazenados nas tabelas de forma não organizadas. Para ordenar, organizar e filtras as informações são necessários os Indices.

BANCOS DE DADOS

MODELAGEM LÓGICA

Normalização é um processo a partir do qual se aplicam regras a todas as tabelas do banco de dados com o objetivo de evitar falhas no projeto, como redundância de dados e mistura de diferentes temas e assuntos numa mesma entidade ou tabela. Durante o projeto de um banco de dados, partir de de um modelo conceitual construirmos o modelo relacional seguindo as boas práticas corretamente porém nem sempre os modelos com os quais iremos nos deparar estão corretamente ajustados para um bom funcionamento.

BANCOS DE DADOS

MODELAGEM LÓGICA

1FN - Primeira Forma Normal: todos os atributos de uma tabela devem ser atômicos, ou seja, a tabela não deve conter grupos repetidos e nem atributos com mais de um valor. Para deixar nesta forma normal, é preciso identificar a chave primária da tabela, identificar a(s) coluna(s) que tem(êm) dados repetidos e removê-la(s), criar uma nova tabela com a chave primária para armazenar o dado repetido e, por fim, criar uma relação entre a tabela principal e a tabela secundária.

Por exemplo, considere a tabela Pessoas a seguir.

PESSOAS = {CPF+ NOME + ENDERECO + TELEFONES}

Ela contém a chave primária natural CPF e o atributo TELEFONES é um atributo multivalorado e, portanto, a tabela não está na 1FN. Para deixá-la na 1FN, vamos criar uma nova tabela chamada TELEFONES que conterá PESSOA_ID como chave estrangeira de PESSOAS e TELEFONE como o valor multivalorado que será armazenado.

PESSOAS = { CPF + NOME + ENDERECO }
TELEFONES = { PESSOA_CPF + TELEFONE }

BANCOS DE DADOS

MODELAGEM LÓGICA

2FN - Segunda Forma Normal: Para estar na 2FN é preciso obrigatoriamente estar na 1FN. Além disso, todos os atributos não chaves da tabela devem depender unicamente da chave primária (não podendo depender apenas de parte dela). Para deixar na segunda forma normal, é preciso identificar as colunas que não são funcionalmente dependentes da chave primária da tabela e, em seguida, remover essa coluna da tabela principal e criar uma nova tabela com esses dados. Por exemplo, considere a tabela ALUNOS_CURSOS a seguir.

ALUNOS_CURSOS = { MATR_ALUNO + COD_CURSO + NOTA + DESCR_CURSO }

Nessa tabela, o atributo DESCR_CURSO depende apenas da chave primária ID_CURSO. Dessa forma, a tabela não está na 2FN.

Para tanto, cria-se uma nova tabela chamada CURSOS que tem como chave primária ID_CURSO e atributo DESCRICAO retirando, assim, o atributo DESCR_CURSO da tabela ALUNOS_CURSOS.

ALUNOS_CURSOS = {MATR_ALUNO + ID_CURSO + NOTA}

CURSOS = {ID_CURSO + DESCRICAO}

BANCOS DE DADOS

MODELAGEM LÓGICA

3FN - Terceira Forma Normal: de igual modo para estar na 3FN o modelo precisa estar na 2FN, que diz que os atributos não chave de uma tabela devem ser mutuamente independentes e dependentes unicamente e exclusivamente da chave primária (um atributo B é funcionalmente dependente de A se, e somente se, para cada valor de A só existe um valor de B).

Para atingir essa forma normal, é preciso identificar as colunas que são funcionalmente dependentes das outras colunas não chave e extraí-las para outra tabela. Considere, como exemplo, a tabela FUNCIONARIOS a seguir.

FUNCIONARIOS = { ID + NOME + ID_CARGO + DESCR_CARGO }

O atributo DESCR_CARGO depende exclusivamente de ID_CARGO (atributo não chave) e, portanto, deve-se criar uma nova tabela com esses atributos. Dessa forma, ficamos com as seguintes tabelas:

FUNCIONARIOS = { ID + NOME + ID_CARGO }
CARGOS = { ID_CARGO + DESCRICAO }

PRÁTICA

EXERCÍCIO 3

CRIAÇÃO DOS MODELOS LÓGICOS

Crie em conjunto com seu time um modelo lógico que implemente a solução descrita no modelo conceitual.

DÚVIDAS