TI MAPEADA · 300 MAPAS MENTAIS DE TI

# Módulo 03

## Banco de Dados, BI e Big Data



30 mapas · mapas 062 a 091

**Neste módulo:**

062 SGBD: Conceitos
063 Modelo Entidade-Relacionamento
064 Cardinalidade
065 Modelo Relacional
066 Chaves
067 Normalização: 1FN, 2FN e 3FN
068 FNBC, 4FN e 5FN
069 Álgebra Relacional
070 SQL: DDL
071 SQL: DML e Consultas
072 SQL: JOINs
073 SQL: Agregação
074 SQL: DCL e TCL
075 Transações e ACID
076 Controle de Concorrência
077 Índices
078 Views, Triggers e Procedures
079 NoSQL: Tipos
080 Teorema CAP e BASE
081 Data Warehouse
082 Modelagem Dimensional
083 ETL e ELT
084 OLTP e OLAP
085 Operações OLAP
086 Data Lake e Lakehouse
087 Big Data: os Vs
088 Hadoop e Spark
089 Data Mining
090 Governança de Dados
091 Qualidade de Dados e Metadados

# SGBD: Conceitos

**O QUE É**

Software que gerencia o armazenamento, o acesso e a integridade dos dados.

## 1 Funções

- controle de acesso
- concorrência
- recuperação
- integridade

## 2 Arquitetura ANSI/SPARC

- externo
- conceitual
- interno

## 3 Independência lógica

- muda o esquema conceitual sem afetar visões

## 4 Independência física

- muda o armazenamento sem afetar o conceitual

## 5 Dicionário de dados

- metadados do banco

**⚠ CAI EM PROVA**

Independência física é mais fácil de obter que a lógica.


Material de estudo e revisão · uso pessoal

# Modelo Entidade-Relacionamento

**O QUE É**

Modelo conceitual de Peter Chen que representa dados com entidades e relações.

## 1 Entidade

- forte
- fraca (depende de outra)

## 2 Atributo

- simples
- composto
- multivalorado
- derivado

## 3 Relacionamento

- liga entidades
- pode ter atributos

## 4 Atributo-chave

- identifica a entidade

## 5 Especialização

- generalização com subtipos

**⚠ CAI EM PROVA**

Entidade fraca é identificada com ajuda da entidade forte dona.


Material de estudo e revisão · uso pessoal

# Cardinalidade

**O QUE É**

Quantas ocorrências de uma entidade se ligam a outra.

## 1 1:1

- um para um

## 2 1:N

- um para muitos

## 3 N:N

- muitos para muitos
- vira tabela associativa

## 4 Mínima

- participação opcional (0) ou obrigatória (1)

## 5 Notação

- Chen
- pé de galinha

**⚠ CAI EM PROVA**

Relacionamento N:N no modelo relacional gera uma nova tabela.

# Modelo Relacional

**O QUE É**

Modelo de Codd que organiza os dados em relações (tabelas).

## 1 Relação

- tabela
- tupla = linha
- atributo = coluna

## 2 Domínio

- conjunto de valores válidos

## 3 Grau

- número de colunas
- cardinalidade: número de linhas

## 4 Integridade de entidade

- chave primária não nula

## 5 Integridade referencial

- chave estrangeira válida

**⚠ CAI EM PROVA**

Em teoria, uma relação não tem tuplas repetidas nem ordem entre linhas.


Material de estudo e revisão · uso pessoal

# Chaves

**O QUE É**

Atributos que identificam ou ligam tuplas.

## 1 Superchave

- identifica de forma única

## 2 Candidata

- superchave mínima

## 3 Primária

- candidata escolhida
- única e não nula

## 4 Alternativa

- candidata não escolhida

## 5 Estrangeira

- referencia a chave de outra tabela

**⚠ CAI EM PROVA**

Toda chave candidata é superchave, mas nem toda superchave é candidata.


Material de estudo e revisão · uso pessoal

# Normalização: 1FN, 2FN e 3FN

**O QUE É**

Processo de reduzir redundância e anomalias decompondo tabelas.

## 1 1FN

- valores atômicos
- sem grupos repetidos

## 2 2FN

- 1FN + sem dependência parcial da chave

## 3 3FN

- 2FN + sem dependência transitiva

## 4 Anomalias

- inserção
- atualização
- exclusão

## 5 Desnormalização

- aceita redundância por desempenho

**⚠ CAI EM PROVA**

Tabela com chave primária simples na 1FN já está na 2FN.


Material de estudo e revisão · uso pessoal

# FNBC, 4FN e 5FN

**O QUE É**

Formas normais avançadas para casos especiais de dependência.

## 1 FNBC

- todo determinante é chave candidata

## 2 4FN

- elimina dependências multivaloradas

## 3 5FN

- elimina dependências de junção

## 4 Relação

- FNBC é mais rígida que a 3FN

## 5 Prática

- a maioria dos projetos vai até a 3FN

**⚠ CAI EM PROVA**

Toda tabela em FNBC está na 3FN, mas o contrário nem sempre.


Material de estudo e revisão · uso pessoal

# Álgebra Relacional

**O QUE É**

Linguagem formal com operações sobre relações que geram novas relações.

## 1 Seleção σ

- filtra linhas

## 2 Projeção π

- escolhe colunas

## 3 Junção

- combina relações

## 4 Conjuntos

- união
- interseção
- diferença

## 5 Outros

- produto cartesiano
- divisão
- renomear ρ

**⚠ CAI EM PROVA**

Seleção atua nas linhas (WHERE); projeção nas colunas (SELECT).

# SQL: DDL

O QUE É

Linguagem de definição de dados: cria e altera a estrutura do banco.

## 1 CREATE

- cria tabela, visão, índice

## 2 ALTER

- altera estrutura

## 3 DROP

- remove o objeto inteiro

## 4 TRUNCATE

- apaga todas as linhas

## 5 Restrições

- PRIMARY KEY
- FOREIGN KEY
- UNIQUE
- CHECK
- NOT NULL

⚠ CAI EM PROVA

**TRUNCATE é DDL em muitos SGBDs e não aceita WHERE.**

# SQL: DML e Consultas

**O QUE É**

Comandos para manipular e consultar os dados.

## 1 INSERT

- insere linhas

## 2 UPDATE

- altera linhas

## 3 DELETE

- remove linhas
- aceita WHERE

## 4 SELECT

- consulta
- FROM
- WHERE
- ORDER BY

## 5 Ordem lógica

- FROM
- WHERE
- GROUP BY
- HAVING
- SELECT
- ORDER BY

**⚠ CAI EM PROVA**

DELETE sem WHERE apaga todas as linhas da tabela.


Material de estudo e revisão · uso pessoal

# SQL: JOINs

O QUE É

Combinam linhas de duas ou mais tabelas por uma condição.

## 1 INNER JOIN

- só linhas com correspondência

## 2 LEFT JOIN

- todas da esquerda
- NULL onde faltar

## 3 RIGHT JOIN

- todas da direita

## 4 FULL OUTER JOIN

- todas dos dois lados

## 5 CROSS JOIN

- produto cartesiano

⚠ CAI EM PROVA

Self join é a tabela juntada com ela mesma.


Material de estudo e revisão · uso pessoal

# SQL: Agregação

**O QUE É**

Funções que resumem grupos de linhas em um valor.

## 1 Funções

- COUNT
- SUM
- AVG
- MIN
- MAX

## 2 GROUP BY

- agrupa linhas

## 3 HAVING

- filtra grupos depois da agregação

## 4 WHERE

- filtra linhas antes da agregação

## 5 COUNT(*)

- conta linhas
- COUNT(coluna) ignora NULL

**⚠ CAI EM PROVA**

Filtro com função de agregação vai no HAVING, não no WHERE.


Material de estudo e revisão · uso pessoal

# SQL: DCL e TCL

**O QUE É**

Comandos de controle de acesso e de transações.

## 1 DCL

- GRANT concede
- REVOKE revoga

## 2 TCL

- COMMIT confirma
- ROLLBACK desfaz

## 3 SAVEPOINT

- ponto de retorno parcial

## 4 Papéis

- ROLE agrupa permissões

## 5 Princípio

- menor privilégio

**⚠ CAI EM PROVA**

ROLLBACK desfaz só o que ainda não recebeu COMMIT.


Material de estudo e revisão · uso pessoal

# Transações e ACID

**O QUE É**

Unidade lógica de trabalho que deve ser executada por completo ou não ser executada.

## 1 Atomicidade

- tudo ou nada

## 2 Consistência

- leva de um estado válido a outro

## 3 Isolamento

- transações não interferem entre si

## 4 Durabilidade

- confirmado não se perde

## 5 Estados

- ativa
- parcialmente confirmada
- confirmada
- falha
- abortada

**⚠ CAI EM PROVA**

Durabilidade é garantida por logs e mecanismos de recuperação.


Material de estudo e revisão · uso pessoal

# Controle de Concorrência

**O QUE É**

Garante o isolamento quando várias transações rodam ao mesmo tempo.

## 1 Problemas

- leitura suja
- leitura não repetível
- fantasma

## 2 Bloqueio

- compartilhado (leitura)
- exclusivo (escrita)

## 3 2PL

- fase de crescimento e fase de encolhimento

## 4 Níveis de isolamento

- read uncommitted até serializable

## 5 Deadlock

- espera circular entre transações

**⚠ CAI EM PROVA**

Serializable é o nível de isolamento mais rígido do padrão SQL.


Material de estudo e revisão · uso pessoal

# Índices

**O QUE É**

Estruturas que aceleram a busca de dados em troca de espaço e escrita mais lenta.

## 1 Árvore B+

- padrão na maioria dos SGBDs

## 2 Hash

- ótimo para igualdade
- ruim para faixas

## 3 Clusterizado

- define a ordem física
- um por tabela

## 4 Não clusterizado

- estrutura separada
- vários por tabela

## 5 Custo

- INSERT e UPDATE ficam mais lentos

**⚠ CAI EM PROVA**

Índice hash não ajuda em consultas por intervalo.


Material de estudo e revisão · uso pessoal

# Views, Triggers e Procedures

**O QUE É**

Objetos programáveis do banco de dados.

## 1 View

- consulta salva
- tabela virtual

## 2 View materializada

- guarda o resultado fisicamente

## 3 Trigger

- dispara em INSERT, UPDATE ou DELETE

## 4 Stored procedure

- rotina armazenada
- chamada explícita

## 5 Function

- retorna valor
- pode ser usada em SELECT

**⚠ CAI EM PROVA**

Trigger dispara automaticamente; procedure precisa ser chamada.


Material de estudo e revisão · uso pessoal

# NoSQL: Tipos

**O QUE É**

Bancos não relacionais pensados para escala e esquemas flexíveis.

## 1 Chave-valor

- Redis
- DynamoDB

## 2 Documento

- MongoDB
- JSON e BSON

## 3 Colunar

- Cassandra
- HBase

## 4 Grafo

- Neo4j
- nós e arestas

## 5 Uso

- grande volume
- alta disponibilidade
- dados variados

**⚠ CAI EM PROVA**

MongoDB é orientado a documentos, não a colunas.


Material de estudo e revisão · uso pessoal

# Teorema CAP e BASE

**O QUE É**

Limites de sistemas distribuídos e o modelo de consistência do NoSQL.

## 1 C

- consistência

## 2 A

- disponibilidade

## 3 P

- tolerância a partição

## 4 CAP

- com partição, escolhe entre C e A

## 5 BASE

- basicamente disponível
- estado flexível
- consistência eventual

**⚠ CAI EM PROVA**

Em sistema distribuído a partição é inevitável; a escolha real é entre C e A.


Material de estudo e revisão · uso pessoal

# Data Warehouse

**O QUE É**

Repositório integrado de dados históricos para apoiar decisões.

## 1 Inmon

- orientado a assunto
- integrado
- não volátil
- variante no tempo

## 2 Abordagem Inmon

- top-down
- DW corporativo normalizado

## 3 Abordagem Kimball

- bottom-up
- data marts dimensionais

## 4 Data mart

- DW departamental

## 5 Staging area

- área de preparação dos dados

**⚠ CAI EM PROVA**

Não volátil: os dados são carregados e consultados, não atualizados como no OLTP.

# Modelagem Dimensional

**O QUE É**

Organiza dados em fatos e dimensões para análises rápidas.

## 1 Fato

- medidas numéricas
- aditivas, semiaditivas ou não aditivas

## 2 Dimensão

- contexto
- quem, onde, quando

## 3 Esquema estrela

- dimensões desnormalizadas

## 4 Floco de neve

- dimensões normalizadas

## 5 Granularidade

- nível de detalhe do fato

**⚠ CAI EM PROVA**

Saldo bancário é medida semiaditiva: não soma ao longo do tempo.


Material de estudo e revisão · uso pessoal

# ETL e ELT

O QUE É

Processos de levar dados das fontes para o ambiente analítico.

## 1 Extração

- coleta das fontes

## 2 Transformação

- limpa
- padroniza
- integra

## 3 Carga

- grava no destino

## 4 ELT

- carrega primeiro e transforma no destino

## 5 Uso do ELT

- data lakes e nuvem

⚠ CAI EM PROVA

No ETL a transformação ocorre antes da carga.


Material de estudo e revisão · uso pessoal

# OLTP e OLAP

**O QUE É**

Sistemas transacionais e sistemas analíticos.

## 1 OLTP

- operações do dia a dia
- muitas transações curtas

## 2 OLTP dados

- atuais
- normalizados

## 3 OLAP

- análise
- consultas complexas

## 4 OLAP dados

- históricos
- desnormalizados

## 5 Tipos OLAP

- MOLAP
- ROLAP
- HOLAP

**⚠ CAI EM PROVA**

ROLAP usa banco relacional; MOLAP usa cubo multidimensional.

# Operações OLAP

O QUE É

Formas de navegar nos cubos de dados.

## 1 Drill down

- mais detalhe

## 2 Roll up

- mais resumo

## 3 Slice

- fixa uma dimensão

## 4 Dice

- subcubo com várias dimensões

## 5 Pivot

- gira os eixos da visualização

⚠ CAI EM PROVA

Drill across consulta fatos diferentes por dimensões em comum.

# Data Lake e Lakehouse

O QUE É

Repositórios para grandes volumes de dados brutos e variados.

## 1 Data lake

- dados brutos
- schema on read

## 2 DW

- dados tratados
- schema on write

## 3 Risco

- data swamp, lago sem governança

## 4 Lakehouse

- lago com recursos de DW

## 5 Camadas

- bronze
- prata
- ouro

⚠ CAI EM PROVA

Data lake aplica o esquema na leitura (schema on read).


Material de estudo e revisão · uso pessoal

# Big Data: os Vs

**O QUE É**

Dados com volume, velocidade ou variedade além do tratamento tradicional.

## 1 Volume

- quantidade muito grande

## 2 Velocidade

- geração e processamento rápidos

## 3 Variedade

- estruturados
- semiestruturados
- não estruturados

## 4 Veracidade

- confiabilidade

## 5 Valor

- utilidade para o negócio

**⚠ CAI EM PROVA**

Os 3 Vs originais são volume, velocidade e variedade.


Material de estudo e revisão · uso pessoal

# Hadoop e Spark

**O QUE É**

Ferramentas de processamento distribuído de grandes volumes de dados.

## 1 HDFS

- sistema de arquivos distribuído
- blocos replicados

## 2 MapReduce

- fases map e reduce
- em disco

## 3 YARN

- gerencia recursos do cluster

## 4 Spark

- processamento em memória
- mais rápido

## 5 Spark módulos

- SQL
- Streaming
- MLlib
- GraphX

**⚠ CAI EM PROVA**

Spark processa em memória; MapReduce grava resultados intermediários em disco.

# Data Mining

**O QUE É**

Descoberta de padrões úteis em grandes bases de dados.

## 1 Tarefas

- classificação
- agrupamento
- associação
- regressão

## 2 Associação

- regras se-então
- algoritmo Apriori

## 3 Métricas de regra

- suporte
- confiança
- lift

## 4 Processo KDD

- seleção
- pré-processamento
- transformação
- mineração
- avaliação

## 5 CRISP-DM

- 6 fases cíclicas

**⚠ CAI EM PROVA**

Data mining é uma etapa do processo KDD.


Material de estudo e revisão · uso pessoal

# Governança de Dados

**O QUE É**

Exercício de autoridade e controle sobre a gestão dos dados.

## 1 DAMA-DMBOK

- referência de gestão de dados

## 2 Áreas

- qualidade
- metadados
- segurança
- arquitetura
- MDM

## 3 Papéis

- data owner
- data steward

## 4 MDM

- dados mestres únicos e confiáveis

## 5 Políticas

- padrões
- responsabilidades
- conformidade

**⚠ CAI EM PROVA**

No DMBOK, a governança de dados fica no centro das demais áreas.

# Qualidade de Dados e Metadados

**O QUE É**

Garantir que os dados sejam adequados ao uso e bem descritos.

## 1 Dimensões

- completude
- precisão
- consistência
- atualidade
- unicidade

## 2 Metadados técnicos

- tipos
- tabelas
- origem

## 3 Metadados de negócio

- definições
- regras

## 4 Linhagem

- caminho do dado da origem ao destino

## 5 Data profiling

- analisa o perfil dos dados

**⚠ CAI EM PROVA**

Metadados são dados que descrevem outros dados.