[MAC0313]
Introdução aos Sistemas de Bancos de Dados

Aula 1
Modelagem e Implementação de Bancos de Dados (Conceitos Introdutórios)

Kelly Rosa Braghetto
(kellyrb@ime.usp.br)

DCC–IME–USP

02 de agosto de 2016

O que é um **banco de dados**?

Você sabe citar exemplos?

O que é um **banco de dados**?

Exemplos

- Uma agenda de telefones e endereços de seus contatos
- O catálogo com as informações do acervo de uma biblioteca
- Os dados de imposto de renda da Receita Federal
- Os registros de matrículas e notas dos alunos de uma universidade
- As informações sobre o estoque e as vendas de uma loja
- Os prontuários dos pacientes de um hospital
- Os registros meteorológicos coletados na cidade de São Paulo
- ...

O que é um **banco de dados**?

Definições:

- **Banco de dados** – coleção de dados relacionados
- **Dados** – fatos conhecidos que podem ser registrados e que possuem significado implícito

Referência: "Sistemas de Bancos de Dados" (6 a edição), Elmasri e Navathe. Pearson, 2010.

O que é um **banco de dados**?

Definições:

- **Banco de dados** – coleção de dados relacionados
- **Dados** – fatos conhecidos que podem ser registrados e que possuem significado implícito

Definições genéricas demais!

Referência: "Sistemas de Bancos de Dados" (6 a edição), Elmasri e Navathe. Pearson, 2010.

Propriedades implícitas de um banco de dados (BD)

1. Representar um aspecto do mundo real (= minimundo)
2. Ser uma **coleção lógica e coerente de dados** com algum **significado inerente**
   Uma coleção "aleatória" de dados não é um BD!
3. Ser projetado, construído e povoado com dados que possuem um objetivo específico
   Um BD deve possuir um grupo de usuários em potencial e algumas aplicações pré-concebidas, nas quais esses usuários estão interessados

Exemplo de BD grande

Facebook (dados de abril de 2014)

- *Data warehouse* com mais de 300 PB (petabytes)
- Diariamente, cerca de 600 TB (terabytes) de novos dados
- Mais de 1 bilhão de usuários ativos
- Grande variedade de aplicações: desde do tradicional processamento em lotes até a análise de grafos (redes), aprendizagem de máquina e análise interativa em tempo real.
- Em 2013, o quantidade de dados armazenados no *data warehouse* **triplicou**

1 petabyte = 1.000 terabytes = 1 quadrilhão de bytes
(∼ 210.000 DVDs)

Fonte:
`https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/`

Softwares para a manutenção de bancos de dados

Um dado BD informatizado pode ser criado e mantido por:

- um programa de aplicação desenvolvido especificamente para essa tarefa

  ou

- um **Sistema de Gerenciamento de Banco de Dados (SGBD)**

  Sistema de software de propósito geral que facilita o processo de **definição**, **construção**, **manipulação** e **compartilhamento** de BDs entre vários usuários e aplicações

Sistema de Gerenciamento de Banco de Dados (SGBD)

Apoia o ciclo de vida de um BD:

- **Definir um BD** $\Rightarrow$ especificar os tipos, as estruturas e as restrições para os dados que serão armazenados no BD
- **Construir um BD** $\Rightarrow$ gravar os dados em algum meio de armazenamento (controlado pelo SGBD)
- **Manipular um BD** $\Rightarrow$ realizar funções como consultas ao BD para recuperar dados específicos, atualizar o BD para refletir mudanças no minimundo, etc.
- **Compartilhar um BD** $\Rightarrow$ permitir que múltiplos usuários e programas acessem-no simultaneamente

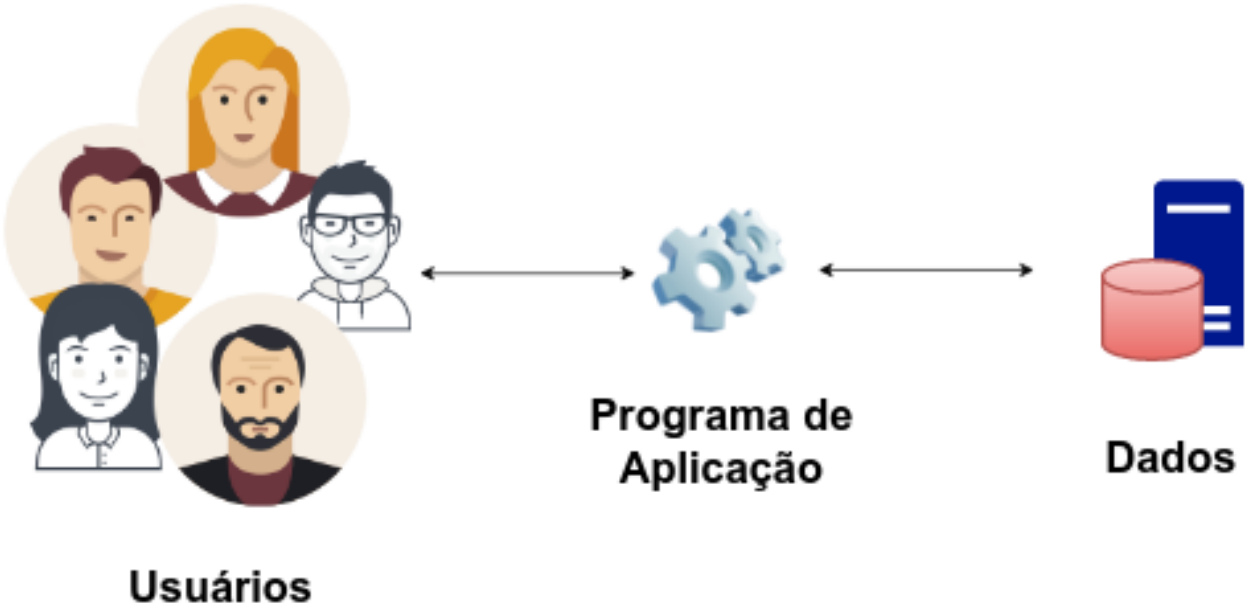
Um sistema de banco de dados monousuário

Um sistema de banco de dados monousuário

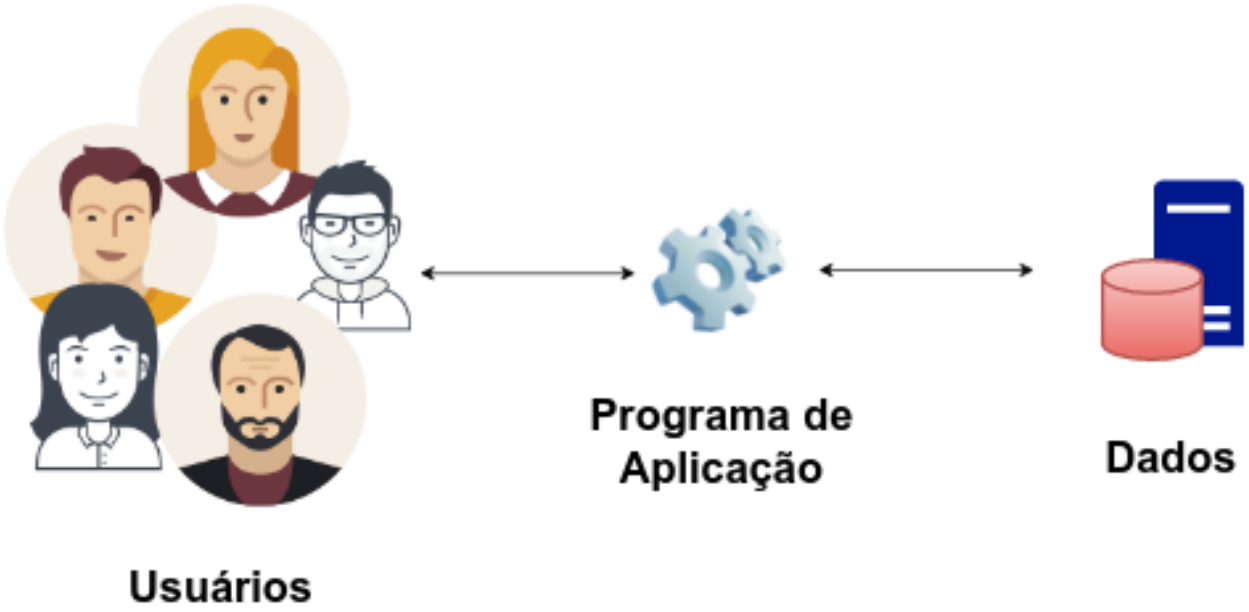
Dificuldades:

- Manter a consistência dos dados
- Possibilitar a recuperação e manipulação eficiente de dados
- Proteger os dados de acessos indevidos
- Recuperar-se de falhas

Um sistema de banco de dados multiusuários



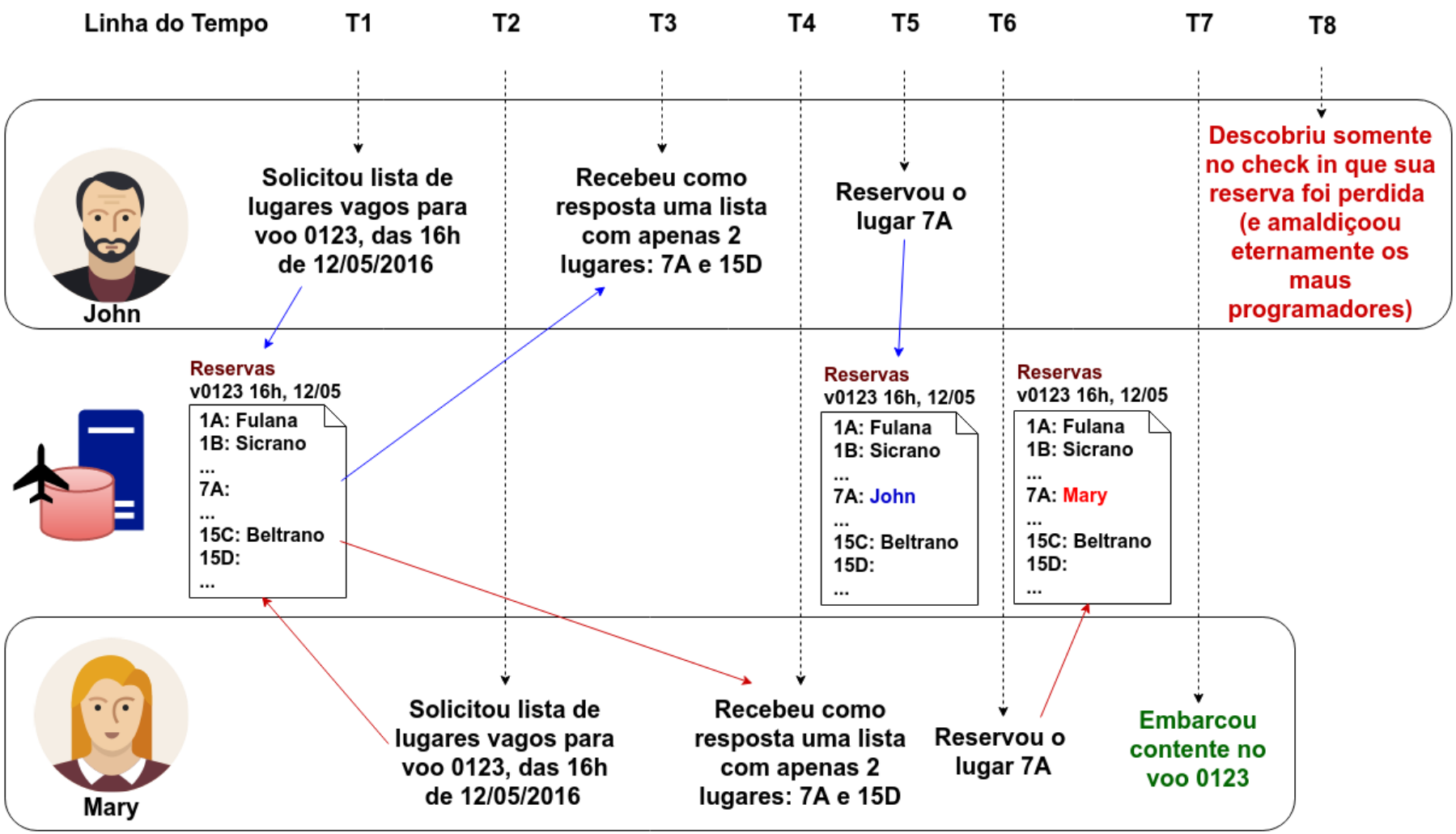
Um sistema de banco de dados multiusuários



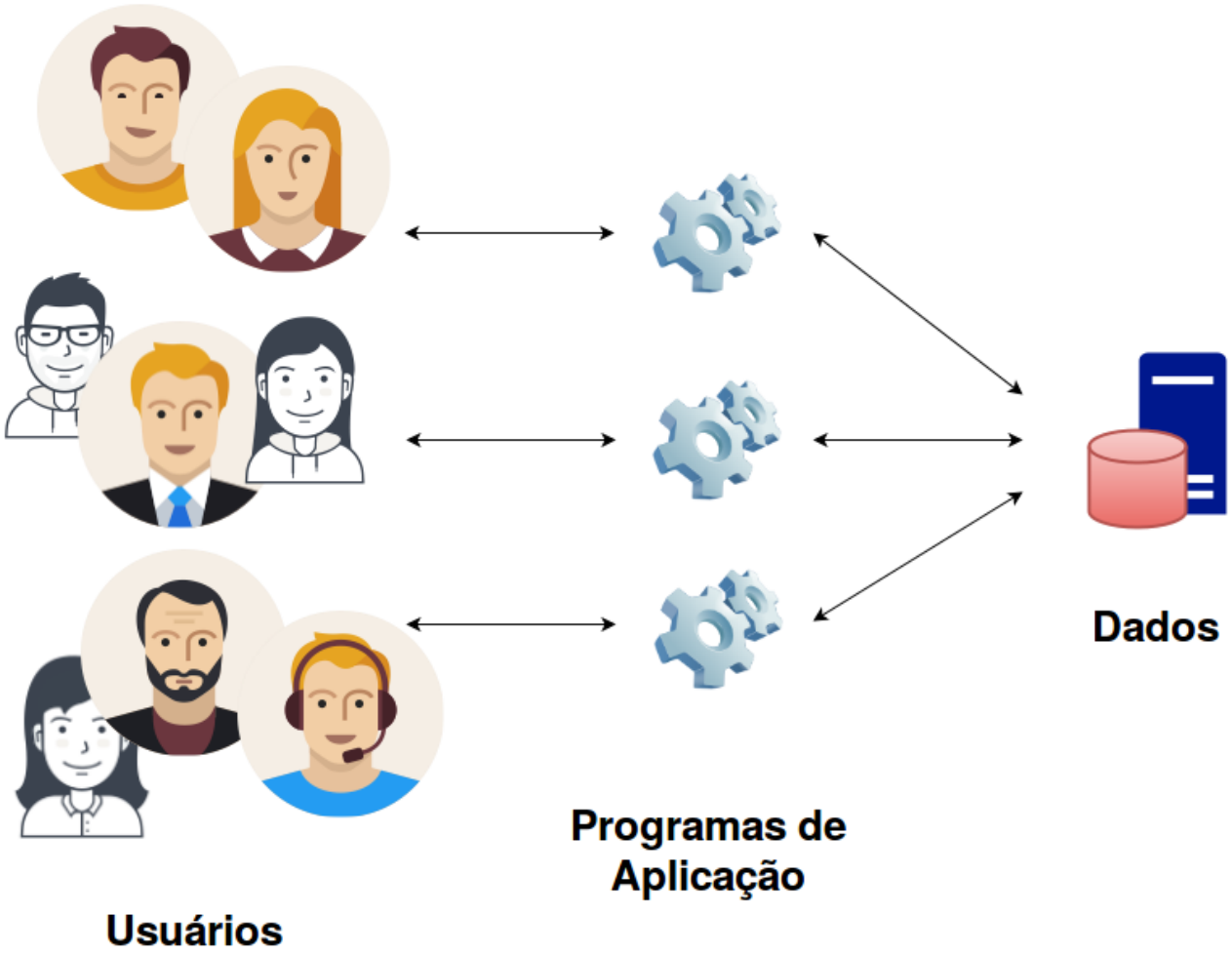
Dificuldades adicionais:

- Gerenciar as necessidades/permissões dos diferentes tipos de usuários
- Evitar conflitos causados por acessos concorrentes aos dados

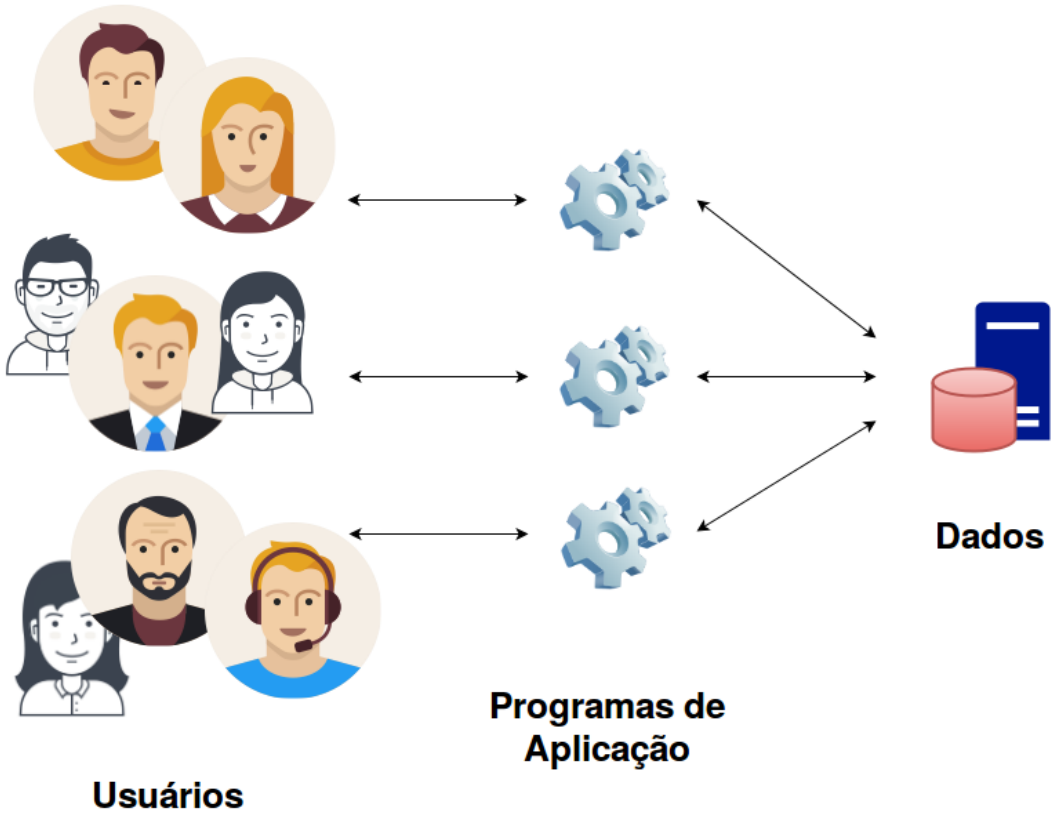
O problema do acesso concorrente aos dados



Outro sistema de banco de dados multiusuários



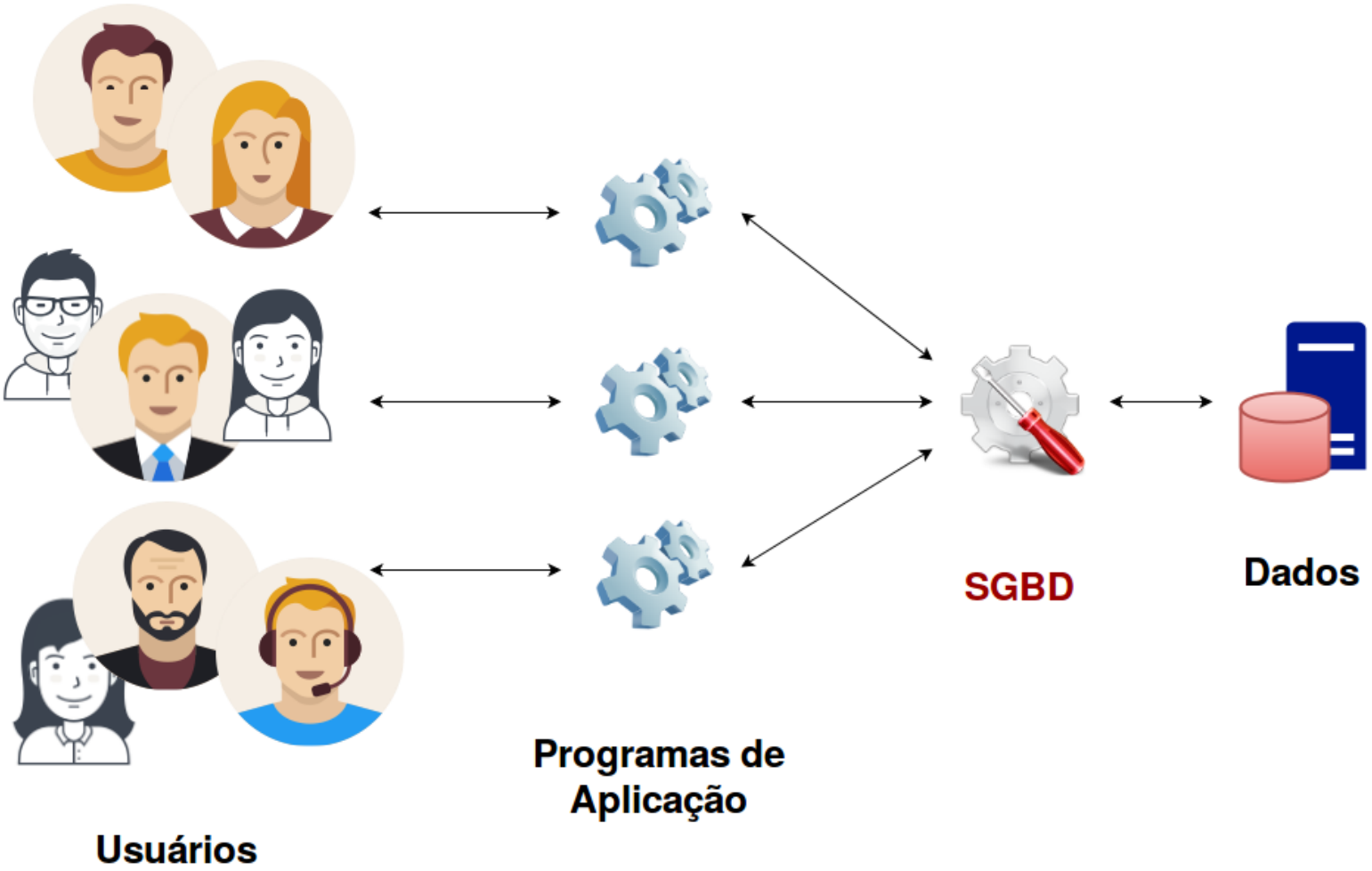
Outro sistema de banco de dados multiusuários



Dificuldades adicionais:

- Manter detalhes sobre a organização dos dados no código de várias aplicações ao mesmo tempo

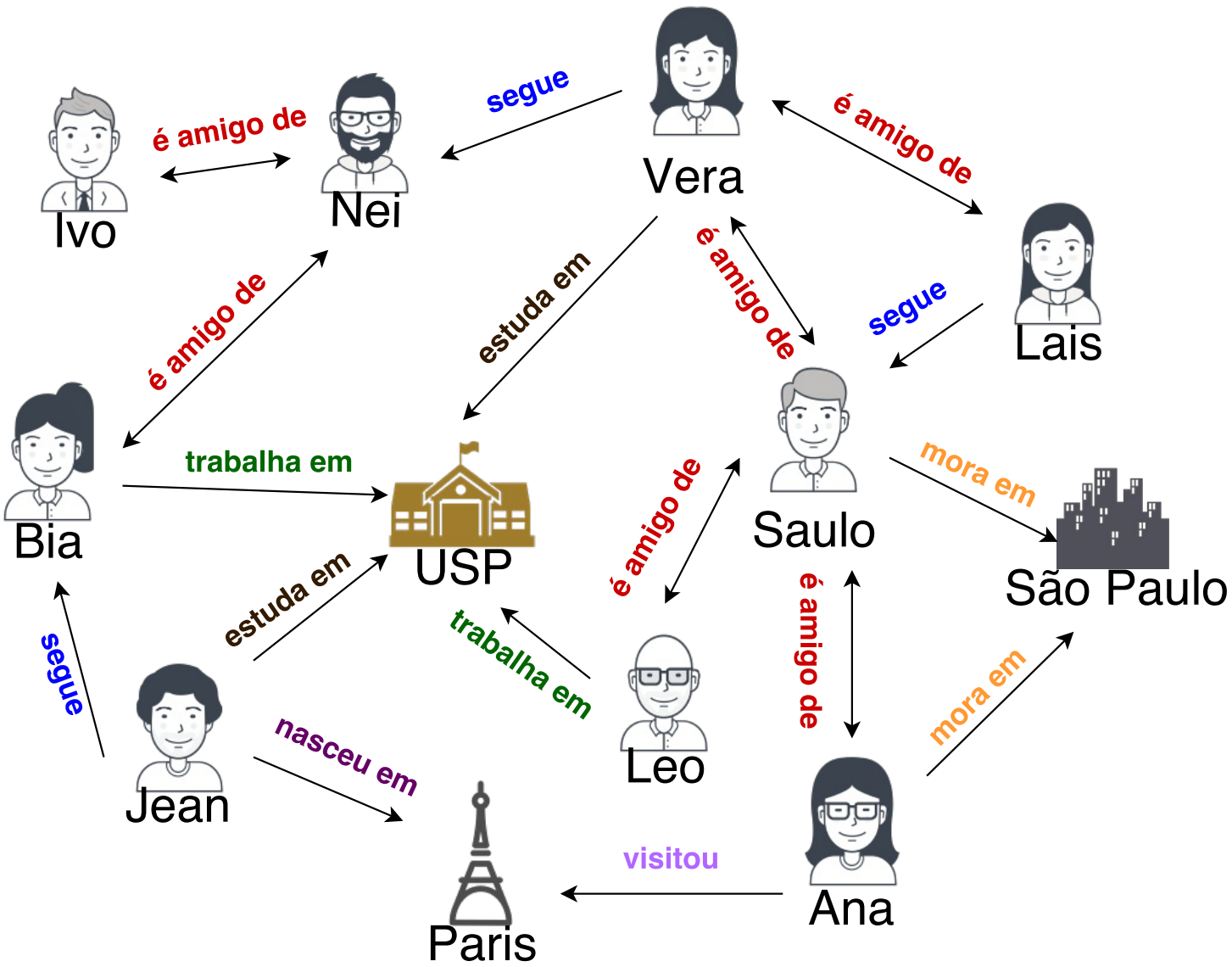
Um sistema de banco de dados com SGBD



Funções Importantes de um SGBD

- Manter os dados por um longo período de tempo
- Garantir a consistência dos dados
- Prover eficiência no acesso aos dados
- Proteger as dados contra acessos não autorizados
- Proteger os dados contra falhas de hardware ou software
- **Prover isolamento entre programas e dados**
- **Prover suporte a visões múltiplas dos dados**
- **Possibilitar o compartilhamento de dados e processar transações multiusuários**

Um exemplo motivacional – dados de uma "rede social"



Um exemplo motivacional – dados de uma "rede social"

Queremos responder perguntas como:

- "Quem são os amigos da Laís?"
- "Quem são os amigos indiretos da Laís?"
- "Existe alguém que não tem nenhum amigo?"
- "Quantos amigos uma pessoa tem em média?"

- De que formas podemos armazenar os dados do grafo em um arquivo em disco?
- Essas formas facilitam a realização das consultas descritas acima?

Representação matricial

	Ana	Bia	Ivo	Jean	Laís	Leo	Nei	Saulo	Vera
Ana								1	
Bia							1		
Ivo							1		
Jean		2							
Laís								2	1
Leo								1	
Nei		1	1						
Saulo	1					1			1
Vera					1		2	1	

1: é amigo de
2: segue

Representação matricial

	Ana	Bia	Ivo	Jean	Laís	Leo	Nei	Saulo	Vera
Ana								1	
Bia							1		
Ivo							1		
Jean		2							
Laís								2	1
Leo								1	
Nei		1	1						
Saulo	1					1			1
Vera					1		2	1	

1: é amigo de
2: segue

Problemas:

- Não é viável para uma rede social com muitas pessoas (desperdiça espaço em disco)
- Se cada linha da matriz corresponde a um registro (ou linha) do arquivo, a inserção de uma nova pessoa na rede social demandaria a reorganização do arquivo todo

Representação com matriz esparsa

Pessoa	Lista de Amigos
Ana	Saulo
Bia	Nei
Ivo	Nei
Jean	
Laís	Vera
Leo	Saulo
Nei	Bia, Ivo
Saulo	Ana, Leo, Vera
Vera	Laís, Saulo

Representação com matriz esparsa

Pessoa	Lista de Amigos
Ana	Saulo
Bia	Nei
Ivo	Nei
Jean	
Laís	Vera
Leo	Saulo
Nei	Bia, Ivo
Saulo	Ana, Leo, Vera
Vera	Laís, Saulo

Problemas:

- Se cada linha da matriz corresponde a um registro (ou linha) do arquivo, a inserção de um novo amigo para uma pessoa já existente na rede social pode demandar a reorganização do arquivo

Representação tabular

Pessoa	Amigo 1	Amigo 2	Amigo 3	Amigo 4	Amigo 5
Ana	Saulo				
Bia	Nei				
Ivo	Nei				
Jean					
Laís	Vera				
Leo	Saulo				
Nei	Bia	Ivo			
Saulo	Ana	Leo	Vera		
Vera	Laís	Saulo			

Representação tabular

Pessoa	Amigo 1	Amigo 2	Amigo 3	Amigo 4	Amigo 5
Ana	Saulo				
Bia	Nei				
Ivo	Nei				
Jean					
Laís	Vera				
Leo	Saulo				
Nei	Bia	Ivo			
Saulo	Ana	Leo	Vera		
Vera	Laís	Saulo			

Problemas:

- A quantidade de amigos por pessoa é limitada

Outra representação tabular

Pessoa	Amigo
Ana	Saulo
Bia	Nei
Ivo	Nei
Laís	Vera
Leo	Saulo
Nei	Bia
Nei	Ivo
Saulo	Ana
Saulo	Leo
Saulo	Vera
Vera	Laís
Vera	Saulo

Outra representação tabular

Pessoa	Amigo
Ana	Saulo
Bia	Nei
Ivo	Nei
Laís	Vera
Leo	Saulo
Nei	Bia
Nei	Ivo
Saulo	Ana
Saulo	Leo
Saulo	Vera
Vera	Laís
Vera	Saulo

Problemas:

- A busca por informações sobre os amigos (diretos e indiretos) de uma pessoa fica mais demorada

Abstração de dados

Oferecer abstração de dados é uma característica fundamental dos SGBDs, ocultando de usuários e aplicações detalhes sobre a organização e armazenamento dos dados

A abstração é feita por meio de modelos de dados

- **Modelo de dados** – conjunto de conceitos usados para descrever a *estrutura* de um banco de dados + *operações* básicas para a recuperação e atualização de dados do banco
- **Estrutura de um banco de dados** – define os tipos de dados, relacionamentos e restrições que se aplicam aos dados

Características de BDs mantidos em SGBDs tradicionais

Natureza autodescritiva

- BDs são mantidos com uma descrição completa de sua estrutura e restrições (metadados)
- Os metadados são armazenados no catálogo do SGBD, e são usados tanto pelo SGBD quanto por usuários do BD
- O SGBD precisa dos metadados porque ele trabalha com qualquer banco de dados (ou seja, é de propósito geral) :))
- Já no processamento de arquivos tradicional, a definição da estrutura dos dados está no código do programa de aplicação $\Rightarrow$ esses programas trabalham com um banco de dados específico :(

Características de BDs mantidos em SGBDs tradicionais

Exemplo: BD com informações de alunos e disciplinas

ALUNO

Nome	Numero_aluno	Tipo_aluno	Curso
Silva	17	1	CC
Braga	8	2	CC

DISCIPLINA

Nome_ disciplina	Numero_ disciplina	Creditos	Departamento
Introd. à ciência da computação	CC1310	4	CC
Estruturas de dados	CC3320	4	CC
Matemática discreta	MAT2410	3	MAT
Banco de dados	CC3380	3	CC

TURMA

Identificacao_ turma	Numero_ disciplina	Semestre	Ano	Professor
85	MAT2410	Segundo	07	Kleber
92	CC1310	Segundo	07	Anderson
102	CC3320	Primeiro	08	Carlos
112	MAT2410	Segundo	08	Chang
119	CC1310	Segundo	08	Anderson
135	CC3380	Segundo	08	Santos

HISTORICO_ESCOLAR

Numero_aluno	Identificacao_turma	Nota
17	112	B
17	119	C
8	85	A
8	92	A
8	102	B
8	135	A

PRE_REQUISITO

Numero_disciplina	Numero_pre_requisito
CC3380	CC3320
CC3380	MAT2410
CC3320	CC1310

Características de BDs mantidos em SGBDs tradicionais

Metadados do BD do slide anterior

RELACOES

Nome_relacao	Numero_de_colunas
ALUNO	4
DISCIPLINA	4
TURMA	5
HISTORICO_ESCOLAR	3
PRE_REQUISITO	2

COLUNAS

Nome_coluna	Tipo_de_dado	Pertence_a_relacao
Nome	Caractere (30)	ALUNO
Numero_aluno	Caractere (4)	ALUNO
Tipo_aluno	Inteiro (1)	ALUNO
Curso	Tipo_curso	ALUNO
Nome_disciplina	Caractere (10)	DISCIPLINA
Numero_disciplina	XXXXNNNN	DISCIPLINA
....	....	....
....	....	....
....	....	....
Numero_pre_requisito	XXXXNNNN	PRE-REQUISITO

Características de BDs mantidos em SGBDs tradicionais

Isolamento entre programas e dados (por meio de **abstração de dados**)

- Um SGBD oferece uma representação conceitual dos dados que não inclui muitos detalhes sobre como eles são armazenados fisicamente
- A inclusão de um novo item de dado na estrutura no BD não implica na necessidade de alteração dos programas de aplicação que o acessam via SGBD; os programas continuarão funcionando corretamente :))
- Já no processamento de arquivos tradicional, qualquer mudança na estrutura de um arquivo implica na necessidade de se alterar todos os programas de aplicação que acessam esse arquivo :(

Características de BDs mantidos em SGBDs tradicionais

Exemplo: formato de armazenamento interno para um registro de ALUNO

RELACOES

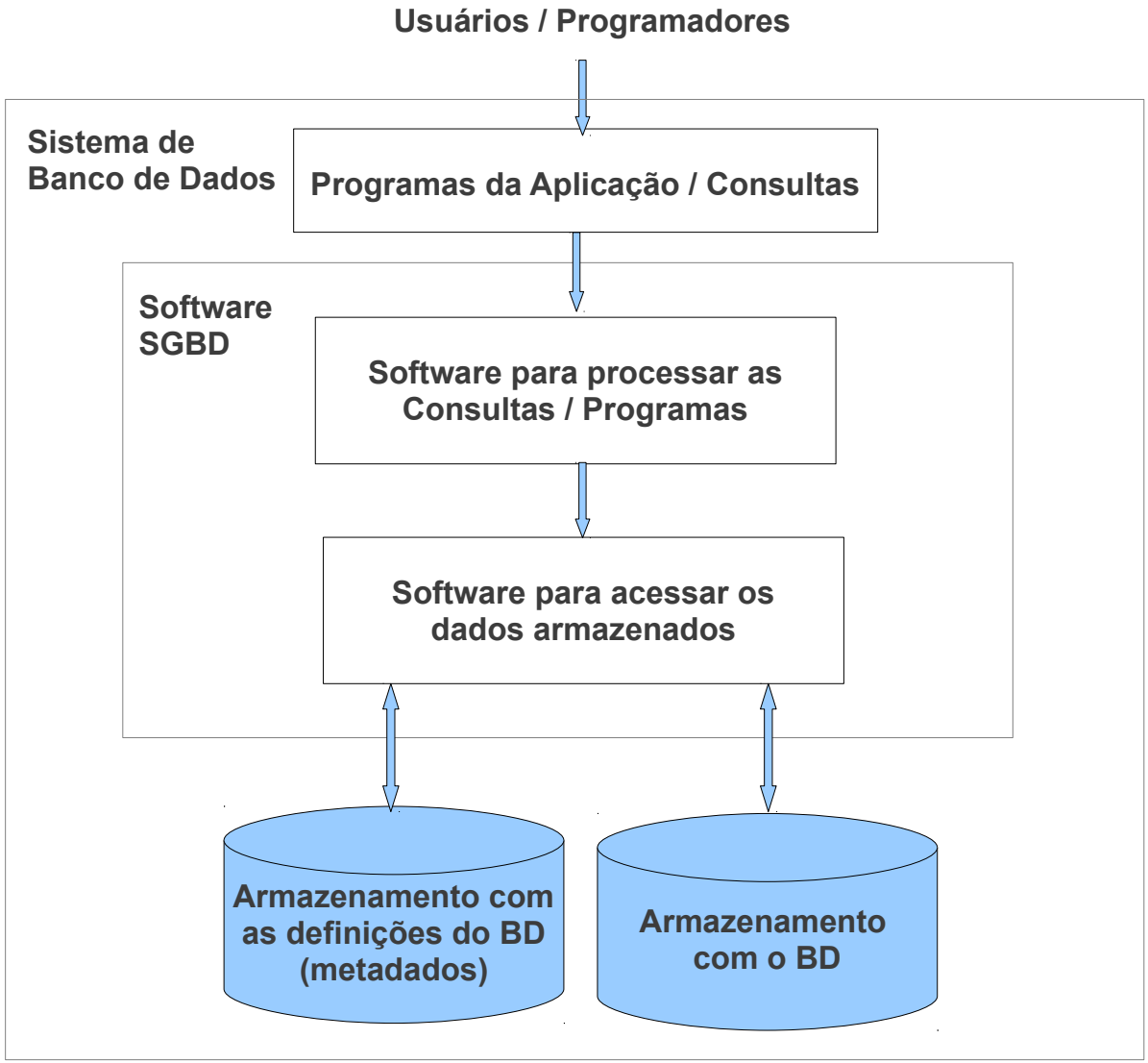
Nome_relacao	Numero_de_colunas
ALUNO	4
DISCIPLINA	4
TURMA	5
HISTORICO_ESCOLAR	3
PRE_REQUISITO	2

COLUNAS

Nome_coluna	Tipo_de_dado	Pertence_a_relacao
Nome	Caractere (30)	ALUNO
Numero_aluno	Caractere (4)	ALUNO
Tipo_aluno	Inteiro (1)	ALUNO
Curso	Tipo_curso	ALUNO
Nome_disciplina	Caractere (10)	DISCIPLINA
Numero_disciplina	XXXXNNNN	DISCIPLINA
....	....	....
....	....	....
....	....	....
Numero_pre_requisito	XXXXNNNN	PRE-REQUISITO

Nome do item de dados	Posicionamento inicial no registro	Tamanho em caracteres (bytes)
Nome	1	30
Numero_aluno	31	4
Tipo_aluno	35	1
Curso	36	4

Sistema de Banco de Dados = Banco de Dados + SGBD



O acesso a um banco de dados

Um usuário ou programa acessa um banco de dados enviando **consultas** ou **transações** para o SGBD que o mantém

- **Consulta** – comando que recupera dados do BD
- **Transação** – comando que lê ou escreve dados do/no BD

Consultas e transações são enviadas a um SGBD por meio de:

- Um software cliente de conexão
- APIs (bibliotecas de funções acessíveis por programação), distribuídas pelos fornecedores do SGBD

Pessoas que interagem com os BDs e seus SGBDs

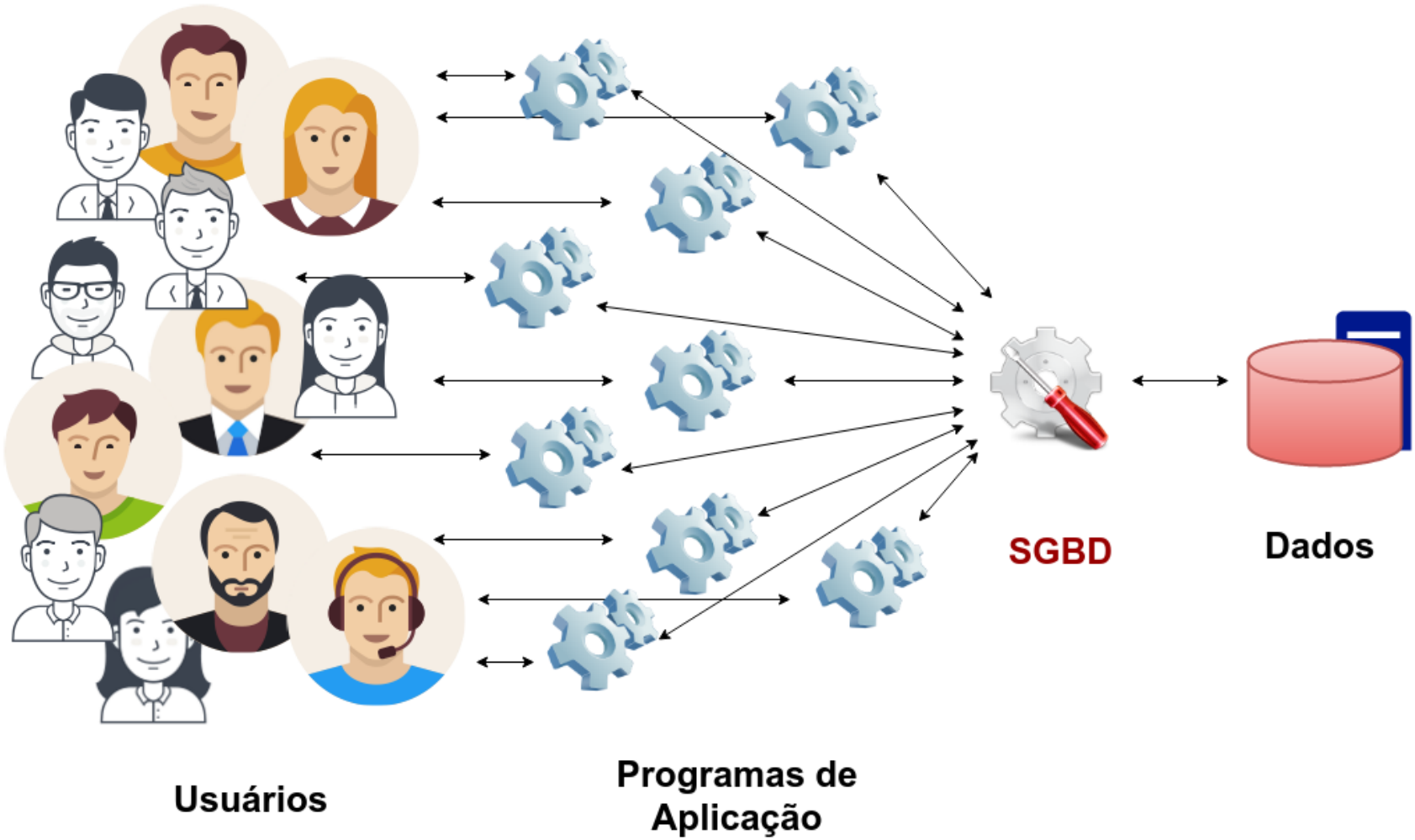
- Administrador de banco de dados (DBA – *database administrator*)
- Projetista de banco de dados
- Usuários finais
  - casuais – usam linguagens de consultas
  - paramétricos – usam transações programadas
  - sofisticados – engenheiros, cientistas, analistas de negócio, ...
  - isolados – mantêm BDs pessoais, usando pacotes prontos
- Analistas de sistemas, programadores, estatísticos, etc.

Linguagens de definição de esquemas e manipulação e consulta de dados

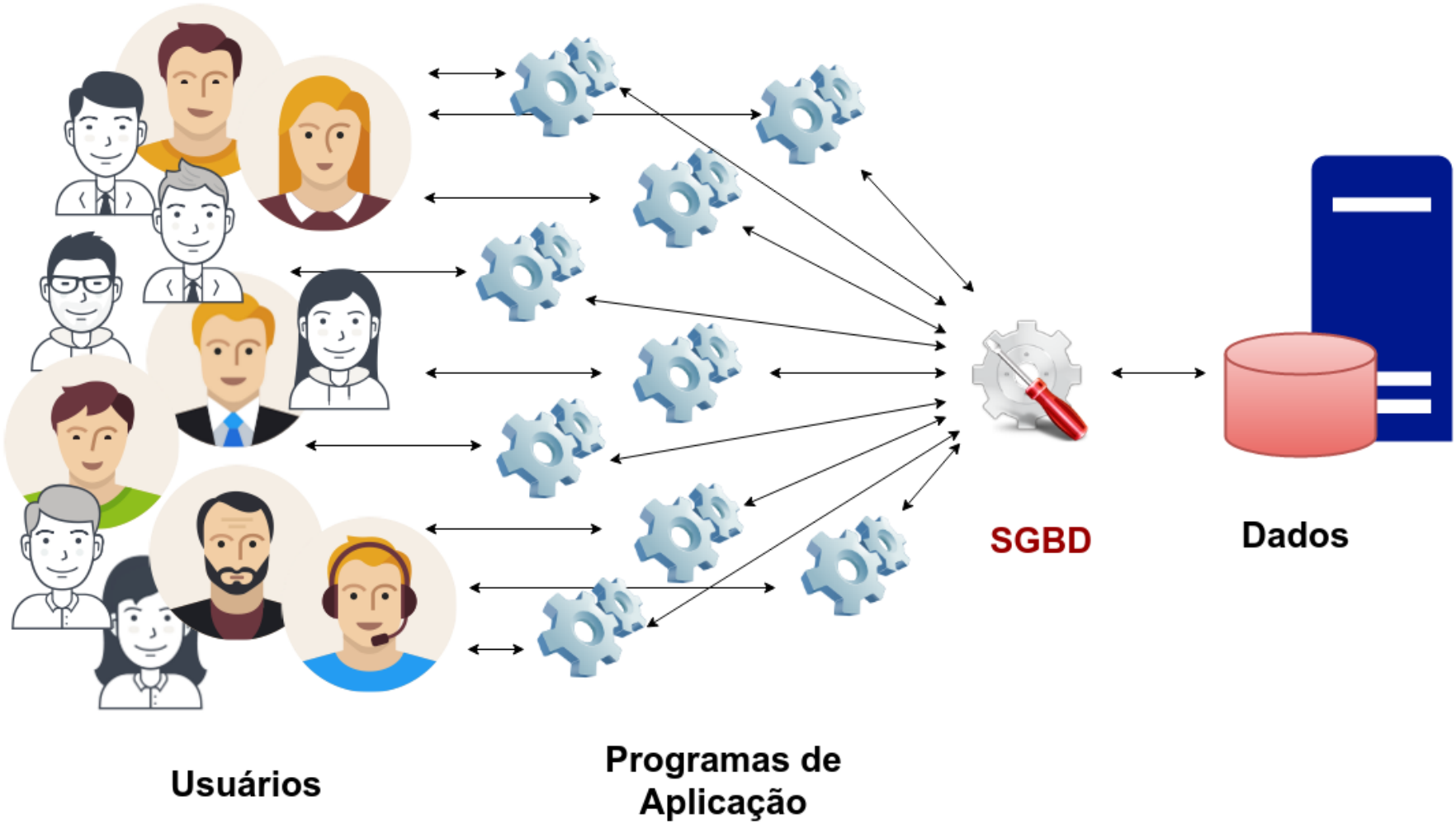
Exemplo de comandos em SQL (*Structured Query Language*)

```
CREATE TABLE Disciplina (
Nome_disciplina     CHAR(30),
Numero_disciplina   CHAR(7),
Creditos            INT,
Departamento        CHAR(3)
);

INSERT INTO Disciplina
VALUES ('Sistemas de Bancos de Dados', 'MAC0426', 4, 'DCC');

SELECT Nome_disciplina, Creditos
FROM   Disciplina
WHERE  Departamento = 'MAT' AND Creditos = 6;
```

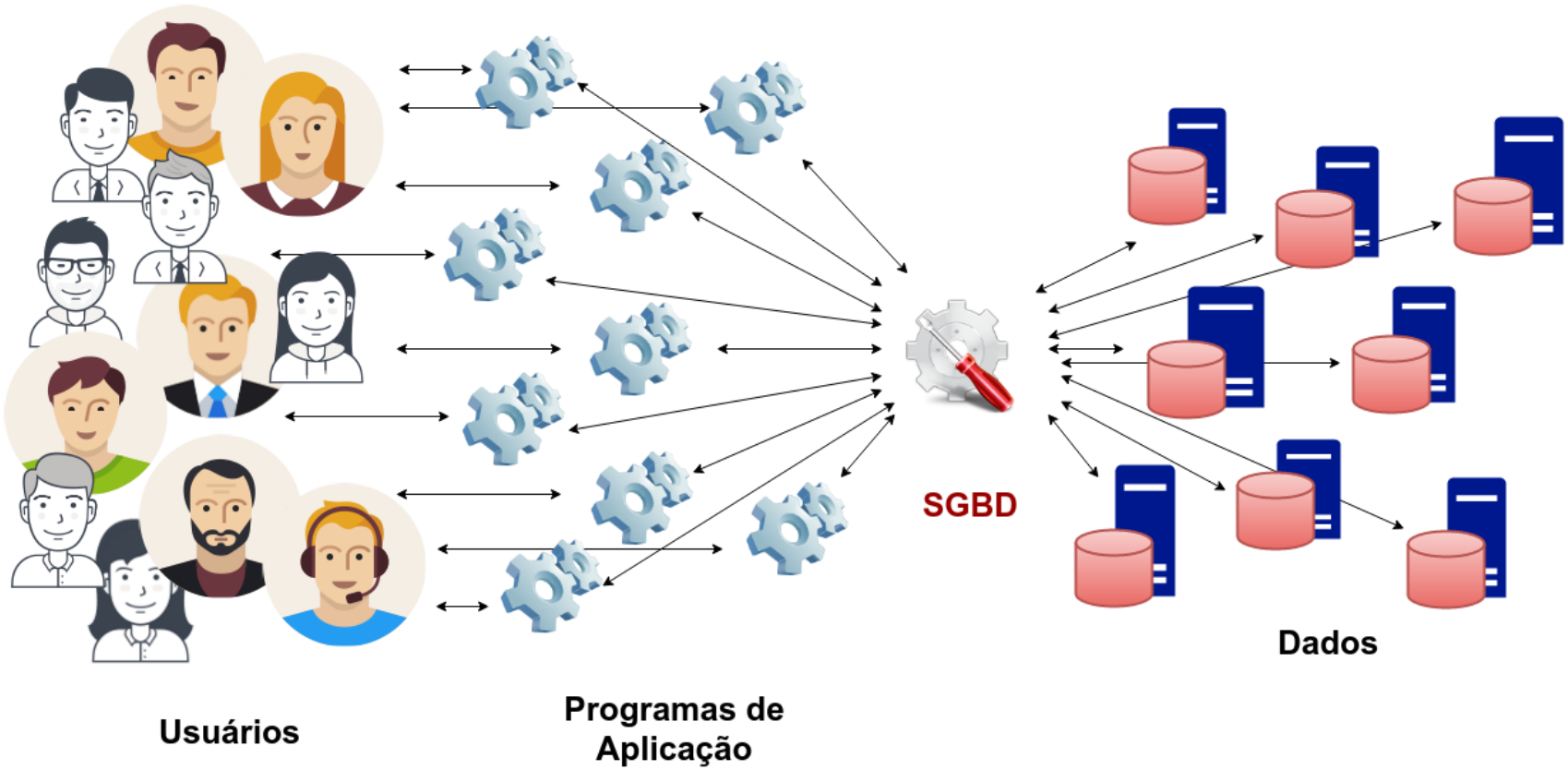
O que fazer quando os dados ou os acessos crescem demais?



SGBD centralizado, escalabilidade vertical



SGBD distribuído, escalabilidade horizontal



Vantagens do uso de um SGBD

1. **Controle de redundâncias** – para evitar *duplicação de esforço*, *desperdício de espaço de armazenamento* e *inconsistência*
2. **Restrição do acesso não autorizado** – por meio da criação de usuários (protegidos por senha) com diferentes tipos de permissões de acesso
3. **Armazenamento persistente para objetos de programas e estruturas de dados**

Vantagens do uso de um SGBD (Cont.)

4. **Estruturas de armazenamento e técnicas de busca para o processamento eficiente de consultas** – por meio da criação de índices e da manutenção de *caches* em memória principal
5. **Mecanismos de *backup* e recuperação** – para garantir a integridade dos dados no caso de falhas
6. **Múltiplas interfaces de usuário** – para apoiar os diferentes perfis de usuários que interagem com os BDs

Vantagens do uso de um SGBD (Cont.)

7. **Capacidade de representação de relacionamentos complexos entre dados**
8. **Imposição de restrições de integridade** – como: de tipo, de integridade referencial, de singularidade (chave)
9. **Possibilidade de deduzir dados e executar ações por meio de regras**

Vantagens do uso de um SGBD (Cont.)

Implicações adicionais:

1. Potencial para garantir padrões
2. Redução no tempo de desenvolvimento de aplicações
3. Flexibilidade
4. Disponibilidade de informações atualizadas
5. Economias de escala

Analisando a viabilidade do uso de um SGBD

Os custos envolvidos no uso de um SGBD se relacionam a:

1. Investimentos iniciais em hardware, software e treinamento [dinheiro, tempo]
2. A generalidade que o SGBD fornece para a definição e o processamento de dados [tempo]
3. O esforço adicional necessário para prover segurança, controle de concorrência, recuperação e integridade dos dados [tempo, desempenho]

Quando é melhor **não** usar um SGBD [convencional]

O uso direto de arquivos ou de SGBDs "não-convencionais" é mais aconselhado que o uso de SGBDs tradicionais (relacionais) nas seguintes situações:

1. O BD e suas aplicações são simples, bem definidos e sem previsão de mudanças
2. A sobrecarga do SGBD pode impedir que requisitos de desempenho (como em programas de tempo-real) sejam atendidos
3. O acesso de múltiplos usuários aos dados não é necessário
4. O dados não se "adequam" ao modelo de dados usado no SGBD

SGBDs mais populares em agosto de 2016

295 systems in ranking, February 2016

Rank Feb 2016	Rank Jan 2016	Rank Feb 2015	DBMS	Database Model	Score Feb 2016	Score Jan 2016	Score Feb 2015
1.	1.	1.	Oracle	Relational DBMS	1476.14	-19.94	+36.42
2.	2.	2.	MySQL	Relational DBMS	1321.13	+21.87	+48.67
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1150.23	+6.16	-27.26
4.	4.	4.	MongoDB	Document store	305.60	-0.43	+38.36
5.	5.	5.	PostgreSQL	Relational DBMS	288.66	+6.26	+26.32
6.	6.	6.	DB2	Relational DBMS	194.48	-1.89	-7.94
7.	7.	7.	Microsoft Access	Relational DBMS	133.08	-0.96	-7.47
8.	8.	8.	Cassandra	Wide column store	131.76	+0.81	+24.68
9.	9.	9.	SQLite	Relational DBMS	106.78	+3.04	+7.22
10.	10.	10.	Redis	Key-value store	102.07	+0.92	+2.86
11.	11.	11.	SAP Adaptive Server	Relational DBMS	80.03	-3.15	-6.30
12.	12.	↑ 16.	Elasticsearch	Search engine	77.84	+0.63	+25.01
13.	↑ 14.	13.	Teradata	Relational DBMS	73.38	-1.57	+3.93
14.	↓ 13.	↓ 12.	Solr	Search engine	72.27	-3.12	-9.21
15.	15.	↑ 17.	Hive	Relational DBMS	52.77	-0.81	+16.21
16.	16.	↓ 14.	HBase	Wide column store	52.02	-1.34	-5.12

http://db-engines.com/en/ranking

Qual é a relação entre Estatística e Bancos de Dados?

- Segundo a Wikipédia [1]:

  A estatística é uma ciência que se dedica à coleta, análise e interpretação de ***dados****.*

- Ferramentas de software estatísticas (também chamadas de "pacotes estatísticos") focam a análise dos dados
- Sistemas de bancos de dados focam o armazenamento e a manipulação dos dados, levando em conta sua estrutura
  - Com isso, eles amparam a coleta e interpretação dos dados

[1] `http://pt.wikipedia.org/wiki/Estatística`)

Sobre os pacotes estatísticos ...

- Geralmente disponibilizam procedimentos que recebem com parâmetro de entrada vetores (*arrays*) de dados e que geram saídas na forma de gráficos, tabelas ou outros vetores de dados

Deficiências comuns desse tipo de software:

- Não manter os dados junto de suas respectivas informações estruturais
- Ser susceptível a dados de má qualidade, como dados incompletos ou inconsistentes
- Não ser capaz de lidar com grandes volumes de dados

Referências Bibliográficas

- *Sistemas de Bancos de Dados* (6ª edição), Elmasri e Navathe. Pearson, 2010.
  Capítulo 1
- *Database Systems – the complete book* (2ª edição), Garcia-Molina, Ullman e Widom. Prentice Hall, 2009.
  Capítulo 1
- *Introdução a Sistemas de Bancos de Dados* (8ª edição), Date. Campus, 2004.
  Capítulos 1 e 2