

# Megadados

Apresentação do curso, introdução a  
sistemas de gerenciamento de bancos de dados

Fábio Ayres <[fabioja@insper.edu.br](mailto:fabioja@insper.edu.br)>



# Bem vindos!

Fábio Ayres

[fabioja@insper.edu.br](mailto:fabioja@insper.edu.br)

Aulas:

- Segundas e quartas, 7:30 – 9:30

Atendimento:

- Segundas, 12:00 – 13:30

# Objetivos de aprendizado

- Entender o que são megadados (big data) e quais os desafios inerentes a dados com esta escala, complexidade, e requisitos de performance
- Dado um problema, estabelecer uma estratégia de trabalho com megadados (integração, armazenamento, processamento, tomada de decisões)
- Projetar software analítico capaz de utilizar estratégias de computação distribuída para tratar de forma eficaz grandes volumes de dados
- Aplicar técnicas de recuperação de informação e mineração de dados.
- Descobrir e avaliar criticamente, de forma autônoma, tecnologias emergentes em big data.

# Estrutura do curso

- Parte I: Bancos de dados relacionais
  - Modelagem
  - SQL
  - Sistemas
- Parte II: Dados em larga escala
  - NoSQL
  - Processamento em lote: MapReduce e Spark
  - Máquinas de busca e recuperação de informação

# Instrumentos de avaliação

## Projetos:

- (APS1) Projeto 1: banco de dados relacional
- (APS2) Projeto 2: big data
- Nota projetos:  $(APS1 + APS2) / 2$

## Provas:

- (P1) Avaliação intermediária
- (P2) Avaliação final
- Nota provas:  $(P1 + P2) / 2$

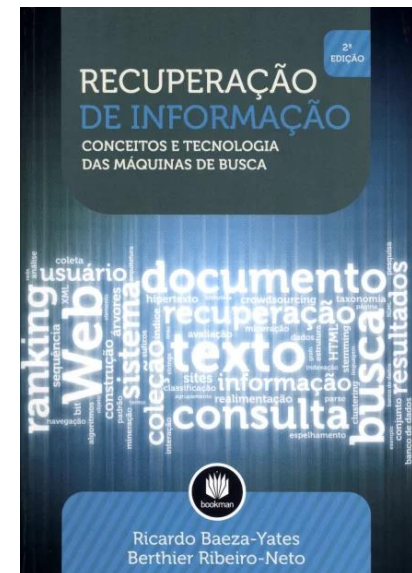
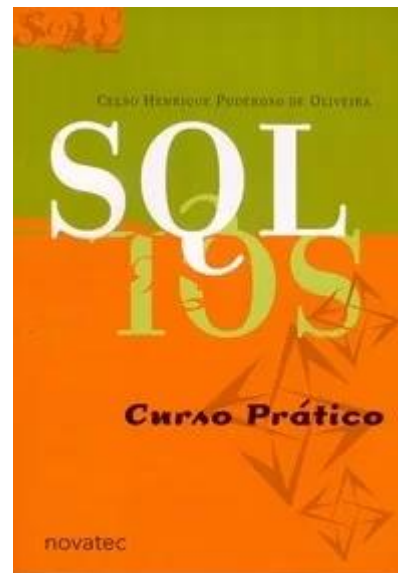
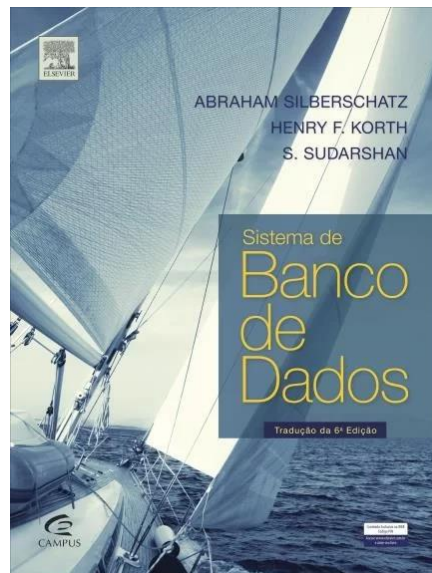
## Nota final:

- Se nota provas e nota projetos  $\geq 5$ : média provas e projetos
- Caso contrário:  $\min(\text{nota provas}, \text{nota projetos})$

# BIBLIOGRAFIA

## BÁSICA

1	SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. <b>Sistema de Banco de Dados</b> . 6a Ed. Rio de Janeiro: Campus, 2012.
2	OLIVEIRA, C. H. P. <b>SQL: Curso Prático</b> . Novatec, 2002.
3	RIBEIRO-NETO, B.; BAEZA-YATES, R. <b>Recuperação de Informação - Conceitos e Tecnologia Das Máquinas de Busca</b> . 2ª Ed. 2013, Bookman



# Para a próxima aula

- Definir grupos para o primeiro projeto

## Instalar

- Anaconda ou alguma versão de Python 3 com Jupyter Notebook
- MySQL Community Server
- MySQL Workbench

# **Sistemas de gerenciamento de bancos de dados**



# Banco de dados: porque?

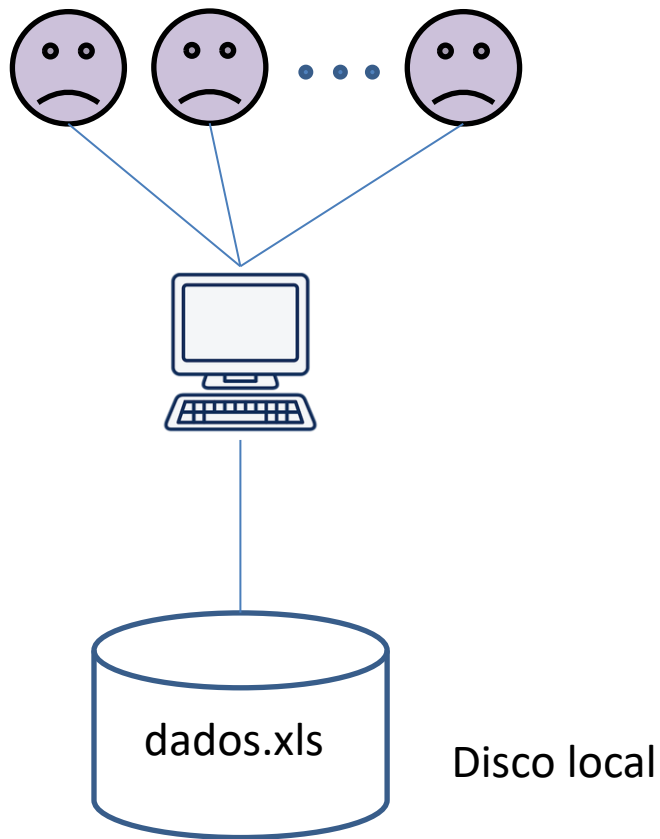
- Você não precisa sempre de um banco de dados para armazenar dados! Você pode usar:
  - Um guardanapo de papel!
  - Um arquivo de texto no seu laptop!
  - Uma planilha Excel!
  - etc...
- Quando será que precisamos de um banco de dados?

# Exemplo

Só um usuário, arquivo pequeno...  
Não precisa de banco de dados!

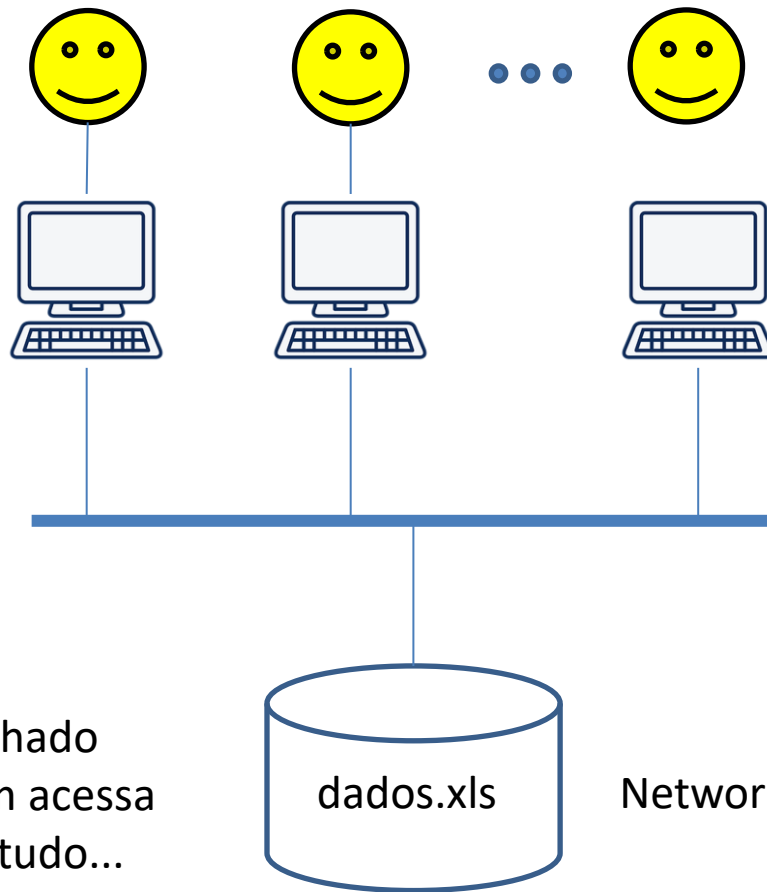


# Exemplo



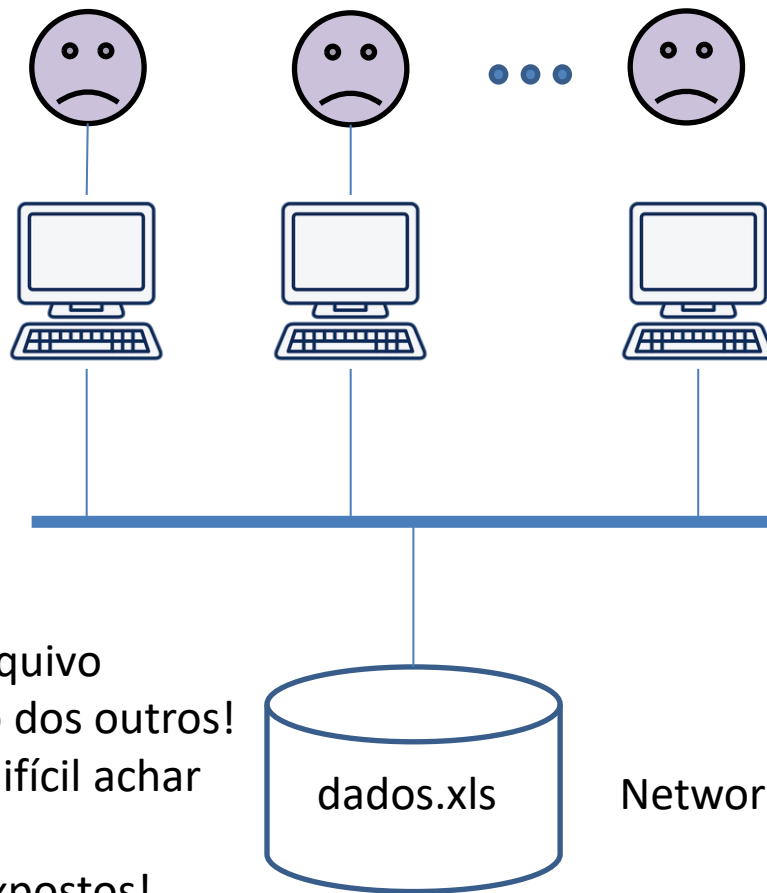
Muitos usuários formando  
fila para conseguir acessar o  
terminal!

# Exemplo



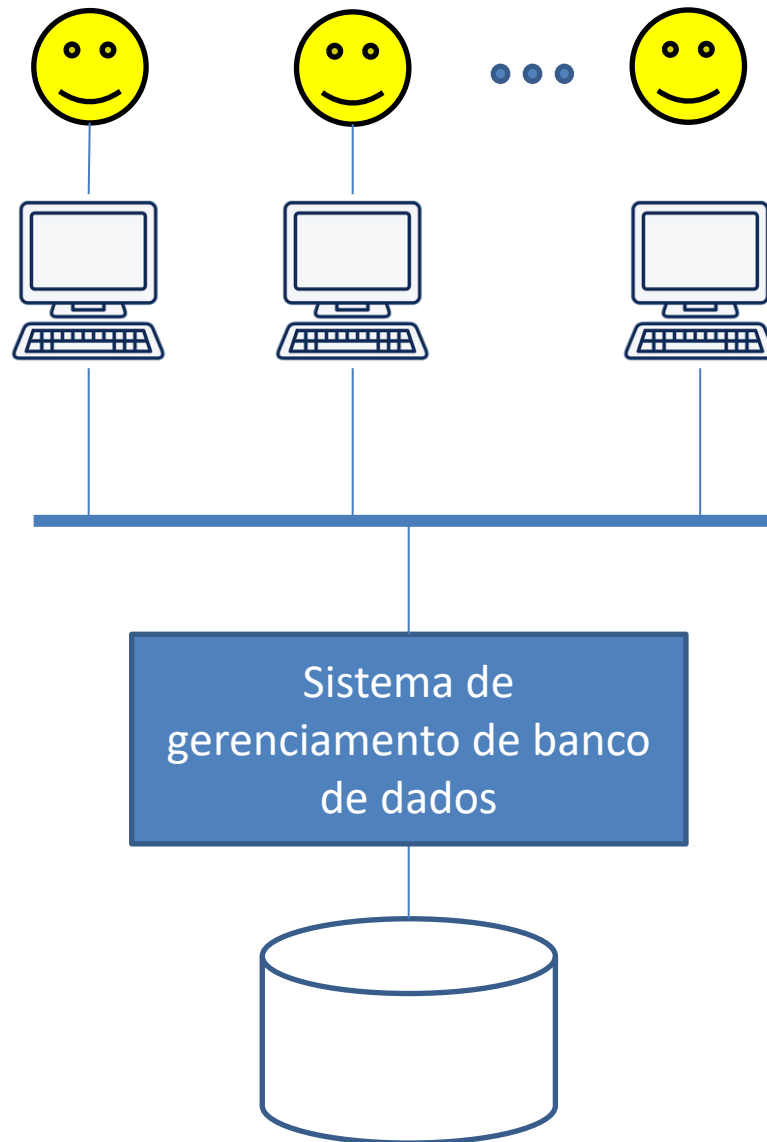
Arquivo está compartilhado na rede agora, cada um acessa com seu terminal. Contudo...

# Exemplo



- Não dá para trabalhar simultaneamente no arquivo sem estragar o trabalho dos outros!
- Dados muito grandes, difícil achar resultados
- Dados sigilosos estão expostos!

# Exemplo



# Motivos para ter um sistema de gerenciamento de banco de dados

- Tamanho
- Velocidade
- Conveniência
- Precisão
- Proteção
- Robustez

Pode não caber na RAM!

Existem maneiras eficientes de armazenar e recuperar dados!

O SGBD já vem com mecanismos sofisticados de consulta!

Um bom projeto evita redundâncias gerencia transações e mantém consistência!

Controle de acesso e registro de atividade!

Proteção contra falhas!

# Aplicações

- Vendas e estoque
- Recursos humanos e informações empresariais
- Dados científicos
- Informações geográficas
- Multimídia
- Jogos online
- Etc!





Sistema de informação

Aplicação/Serviço de consulta



Sistema de Gerenciamento de Banco de Dados

Processamento de consultas

Acesso aos dados



Armazenamento

# Tipos de banco de dados

- Relacional (também chamado de bancos de dados SQL): representa os dados usando o modelo relacional, onde dados são representados através de tabelas bidimensionais.
  - Este é o modelo mais usado em bancos de dados atualmente

# Tipos de banco de dados

- NoSQL: bancos de dados não-relacionais, dentre os quais destacam-se:
  - Key-value stores (e.g. Redis)
  - Document stores (e.g. MongoDB)
  - Column-oriented (e.g. Cassandra)

(Artigo interessante:

<http://www.dataversity.net/review-pros-cons-different-databases-relational-versus-non-relational/>)

# Tipos de banco de dados

- NewSQL: Nova geração de bancos de dados que mesclam as vantagens de alguns tipos de bancos NoSQL (como escalabilidade e disponibilidade) com garantias de consistência transacional do SQL.
  - Exemplo: Google Spanner

(Artigo interessante:

<http://www.odbms.org/blog/2018/03/on-rdbms-nosql-and-newsql-databases-interview-with-john-ryan/>)

# Big data

Atividade:

Fase 1: 15 minutos

- Vamos nos organizar aleatoriamente em grupos de 3 alunos
- Seja  $n$  o numero do seu grupo.  
Responda a pergunta  $((n - 1) \% 3) + 1$  (1 -> 1, 2 -> 2, etc)

Fase 2: 15 minutos

- Todos os grupos da mesma pergunta se juntam e montam uma apresentação de 5 min sobre a sua resposta.

Fase 3: o tempo que for necessário

- Cada grupo faz sua apresentação e discutimos

# Big data

1. Quais são os 5 V's do big data? Explique.
2. Quais os tamanhos típicos de armazenamento de:
  - Um disco SSD
  - Um datacenter
  - Toda a capacidade de armazenamento no mundo
3. A máquina de busca do Google leva tipicamente uma fração de segundo para obter resultados. Como você acha que isso funciona tão rápido?

# Insper

[www.insper.edu.br](http://www.insper.edu.br)