

Megadados

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

macielcv@insper.edu.br
marciofsj@insper.edu.br

Maciel C. Vidal
Marcio Stabile

Bem-vindos!

Maciel C. Vidal

macielcv@insper.edu.br

Marcio F. Stabile Jr.

marciofsj@insper.edu.br

Aulas:

- Segundas e quartas, 7:30 – 9:30

Atendimento:

- Seg, 13:30 – 15:00

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:

- (P1) Projeto 1: banco de dados relacional
- (P2) Projeto 2: big data
- **Nota projetos (NP):**
 - $NP = (P1 + P2) / 2$

Avaliações:

- (AI) Avaliação intermediária
- (AF) Avaliação final
- **Nota avaliações (NA):**
 - $NA = AI*0.4 + AF*0.6$

Atividades:

- Diversas (T1, T2, ..., Tn)
- **Nota atividades (NT):** $SUM(T1, T2, ..., Tn) / n$

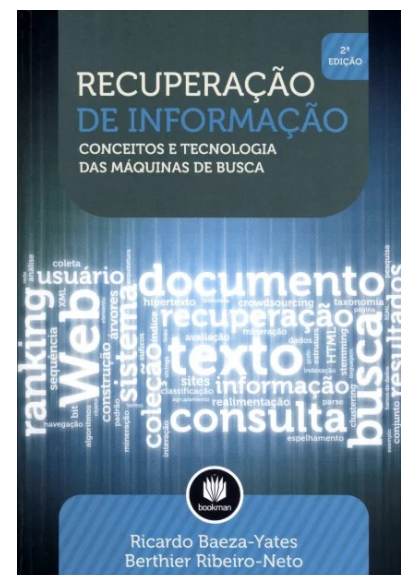
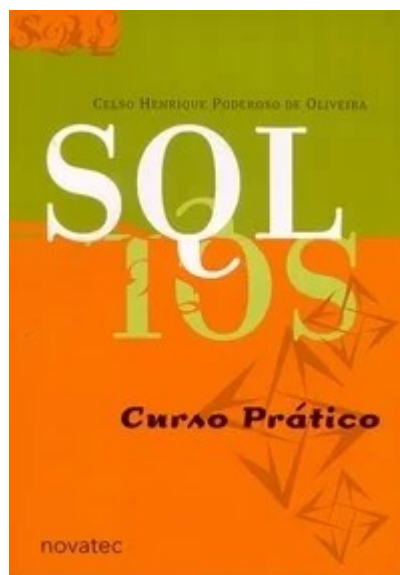
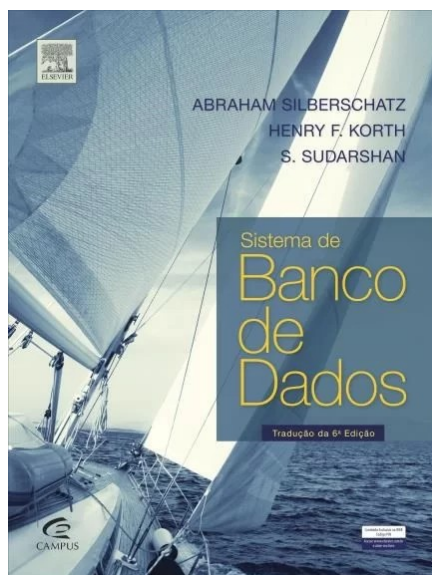
Nota final (NF):

- Se entregou e tirou pelo menos D em todos os projetos
- E nota avaliações, nota projetos e nota atividades ≥ 5 :
 - $NF = NA*0.45 + NP*0.45 + NT*0.1$
- Caso contrário: $NF = \min(NA, P1, P2, NT)$

BIBLIOGRAFIA

BÁSICA

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



Para a próxima aula

- Definir grupos para o primeiro projeto (duplas)

Instalar

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

Banco de dados: porque?

- Quais formas de armazenamento vocês conhecem?!

Sistemas de gerenciamento de bancos de dados SGDB

Banco de dados: porque?

- Quais formas de armazenamento vocês conhecem?!
 - Um guardanapo de papel!
 - Um arquivo de texto no seu laptop!
 - Uma planilha Excel!
 - Etc...

Nem sempre precisamos de um banco de dados para armazenar dados!

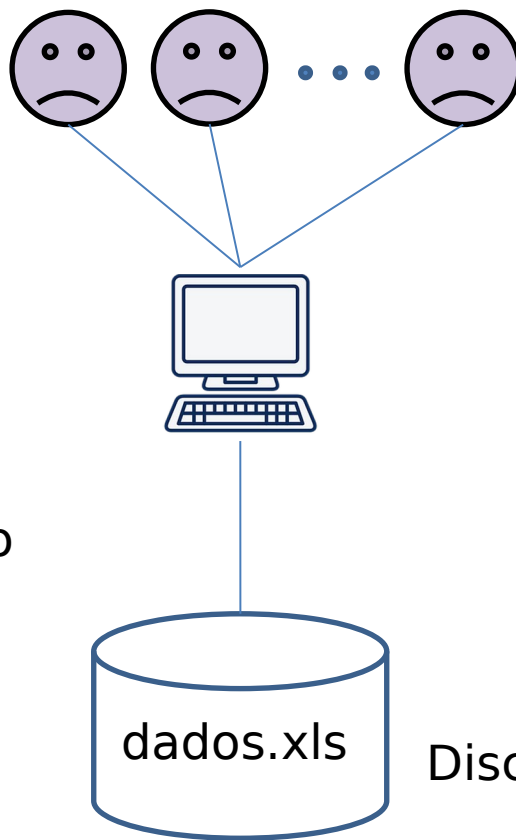
- Quando será que **precisamos** de um **banco de dados**?

Exemplo

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

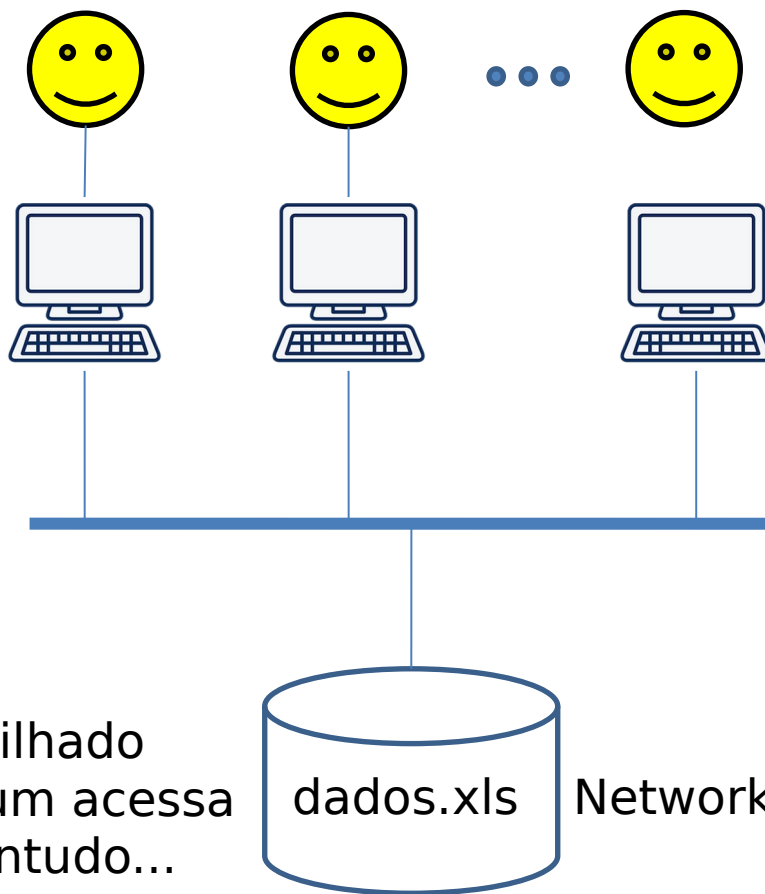


Exemplo



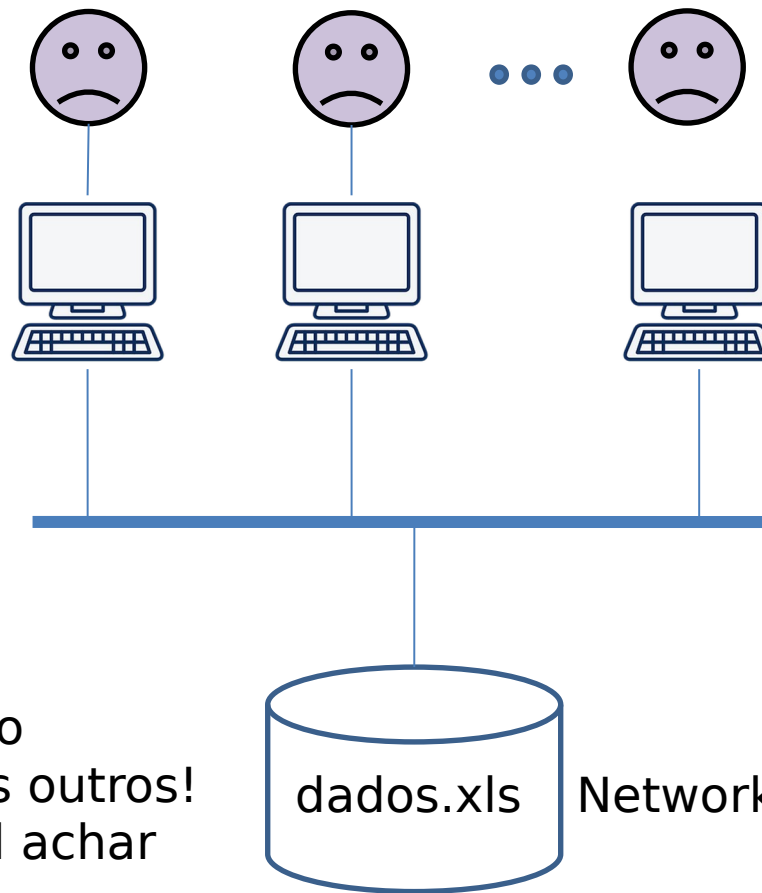
Muitos usuários formando fila para conseguir acessar o terminal!

Exemplo



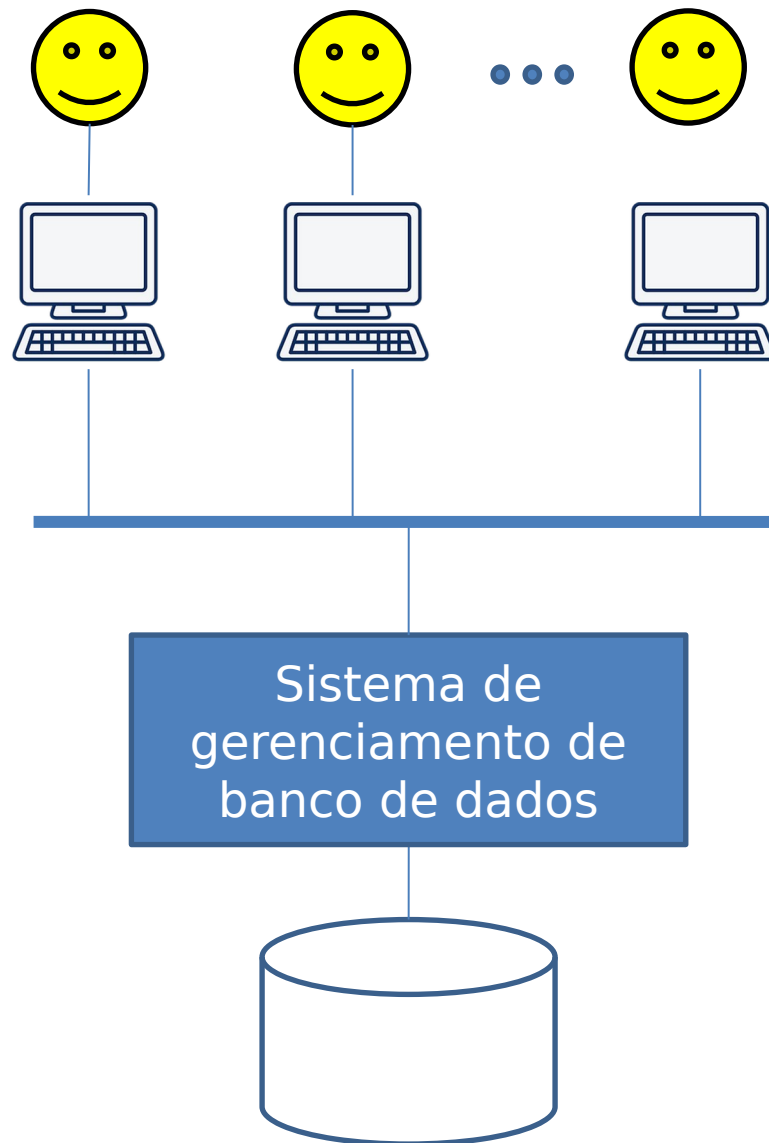
Arquivo está compartilhado na rede agora, cada um acessa com seu terminal. Contudo...

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/>)

Agora...

Instalar

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

Após, abra o notebook da aula para:

- **Instalar pacote autograding**
- **Configurar autograding**

Insper

www.insper.edu.br