

UFSC-CTC-INE-PPGCC

INE 410131 – Gerencia de Dados para Big Data

Aula 2 – Introdução à Big Data

Ronaldo S. Mello

2024/1

Roteiro

1. Timeline dos Modelos de BD
2. Definição de Big Data
3. Os “X”s Vs da Big Data
4. Alguns Domínios de Aplicação
5. Considerações Finais

Roteiro

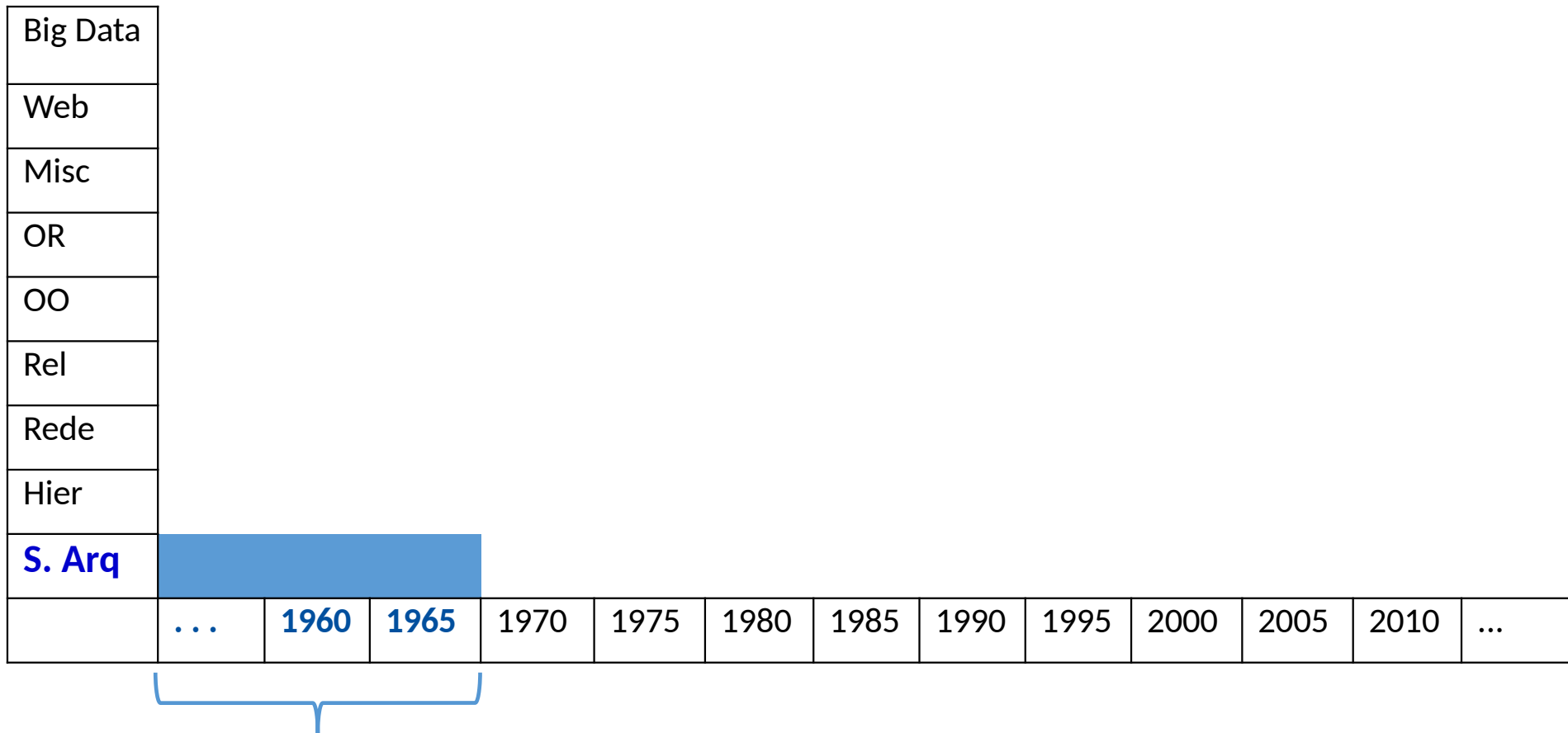
1. Timeline dos Modelos de BD
2. Definição de Big Data
3. Os “X”s Vs da Big Data
4. Alguns Domínios de Aplicação
5. Considerações Finais

Timeline dos Modelos de BD

Big Data												
Web												
Misc												
OR												
OO												
Rel												
Rede												
Hier												
S. Arq												
...	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005	2010	...

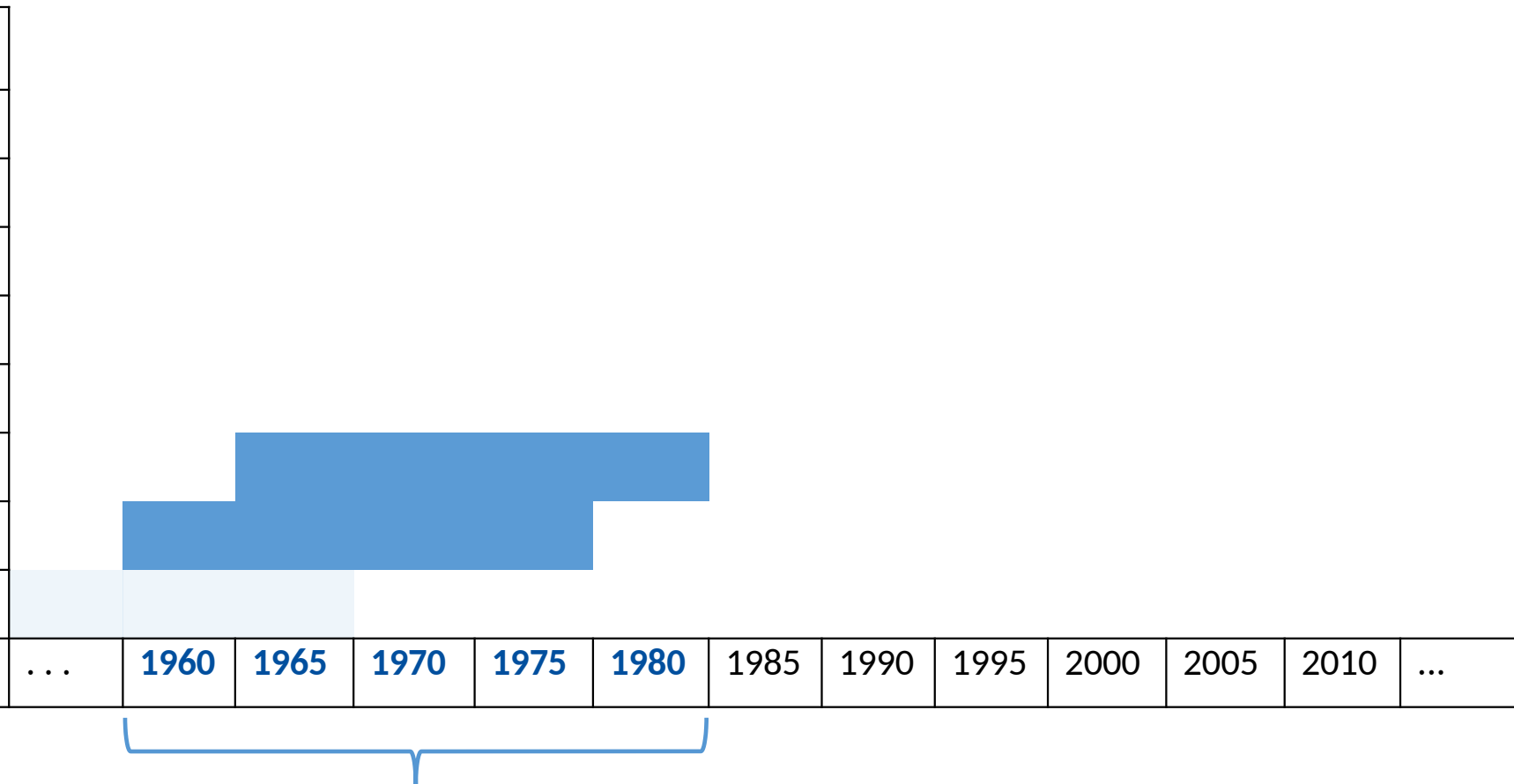
1. Sistemas de Arquivos
2. Modelo Hierárquico
3. Modelo de Rede
4. Modelo Relacional
5. Modelo Orientado a Objetos
6. Modelo Objeto-Relacional
7. Miscelânea (Modelos de dados para propósitos específicos: BD Geográfico, Biológico, Multimídia, ...)
8. Modelos de dados para Web (BDs semiestruturados, XML)
9. Modelos de dados para Big Data (NoSQL, NewSQL, *in-Memory*, ...)

Timeline dos Modelos de BD



- Definição e manipulação de registros simples e fixos
- Gerenciamento apenas de baixo nível dos dados (armazenamento físico)
- Métodos de acesso limitados
- Gerenciamento de integridade e segurança a cargo do aplicação

Timeline dos Modelos de BD



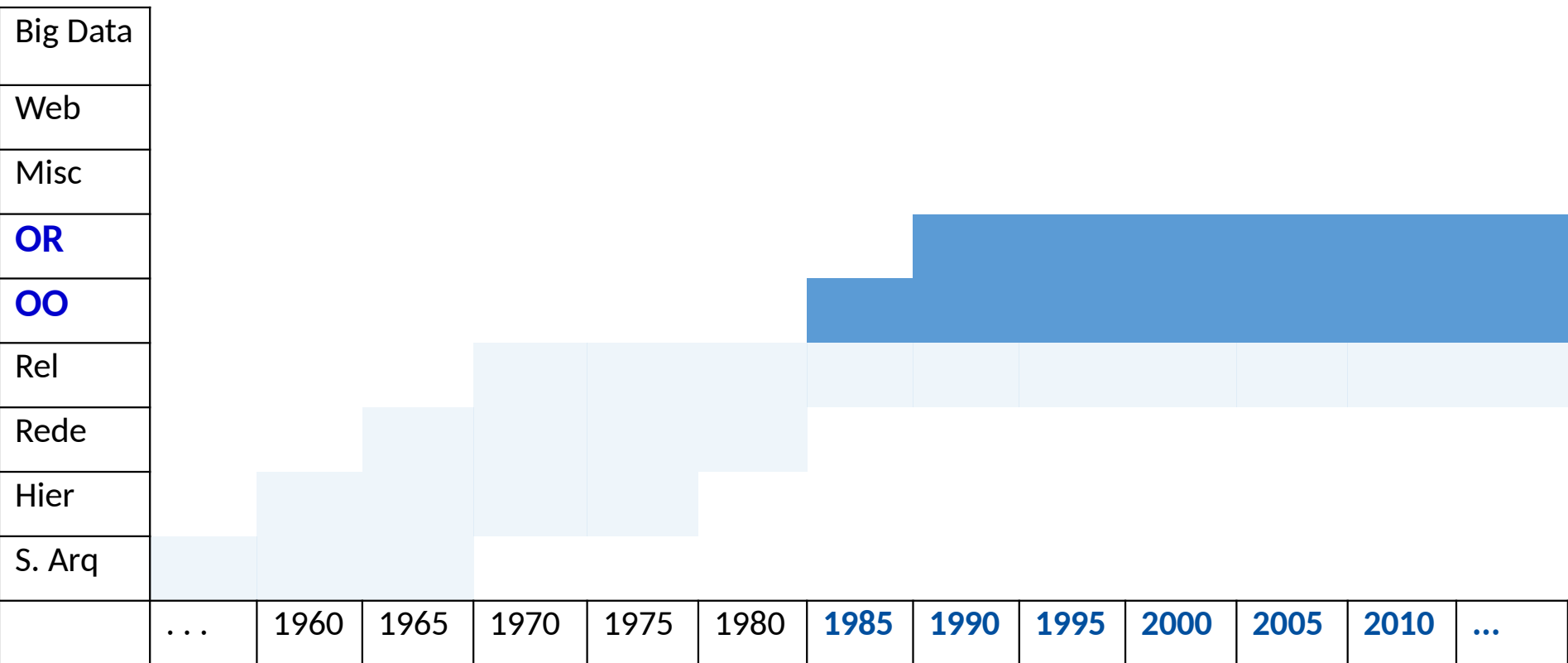
- + Gerenciamento de integridade, concorrência e segurança
- Métodos de acesso limitados
- Modelos de dados limitados (hierarquias, redes de ligações entre registros, ...)

Timeline dos Modelos de BD



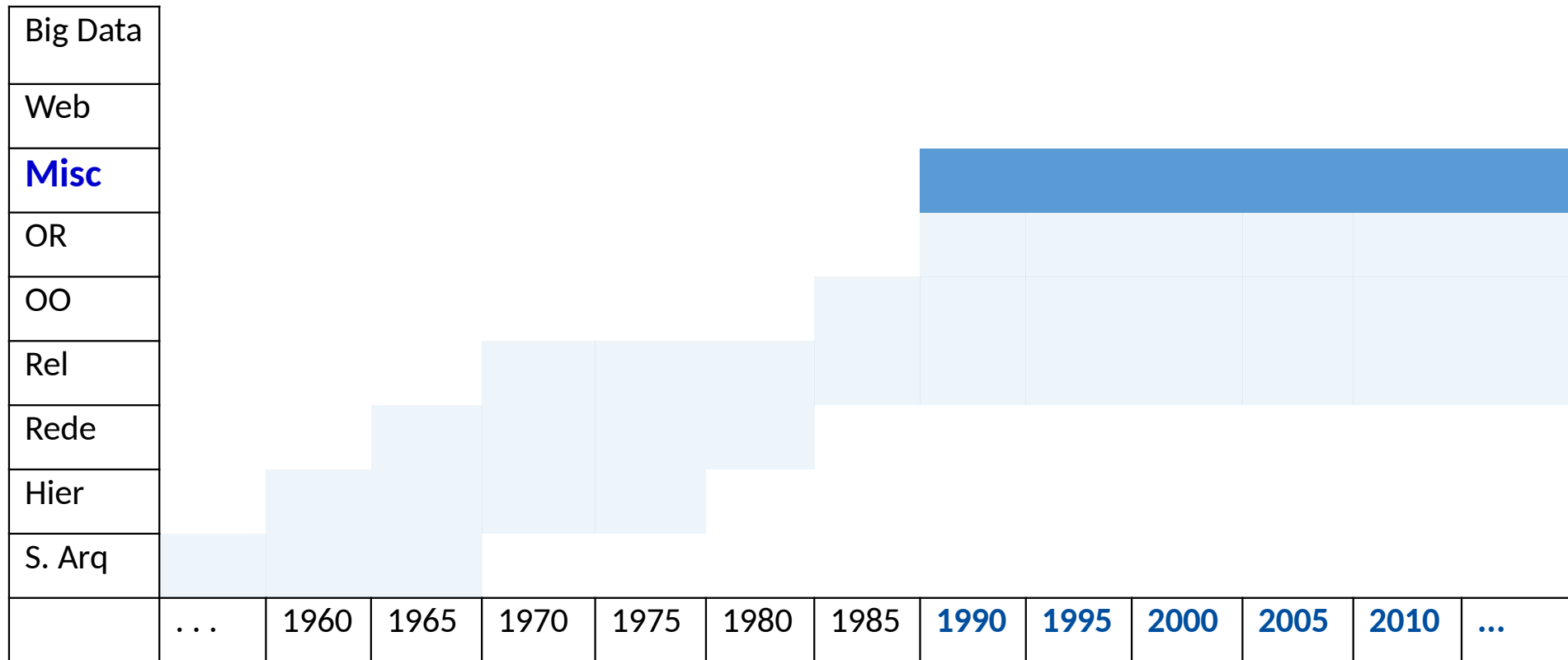
- Linguagens de consulta (flexibilidade de acesso)
- Sólida base formal (teoria de conjuntos – provê otimização de consultas)
- Modelo de dados simples e menos limitado (sem hierarquias, sem restrições para estabelecer ligações (junções) entre registros)

Timeline dos Modelos de BD



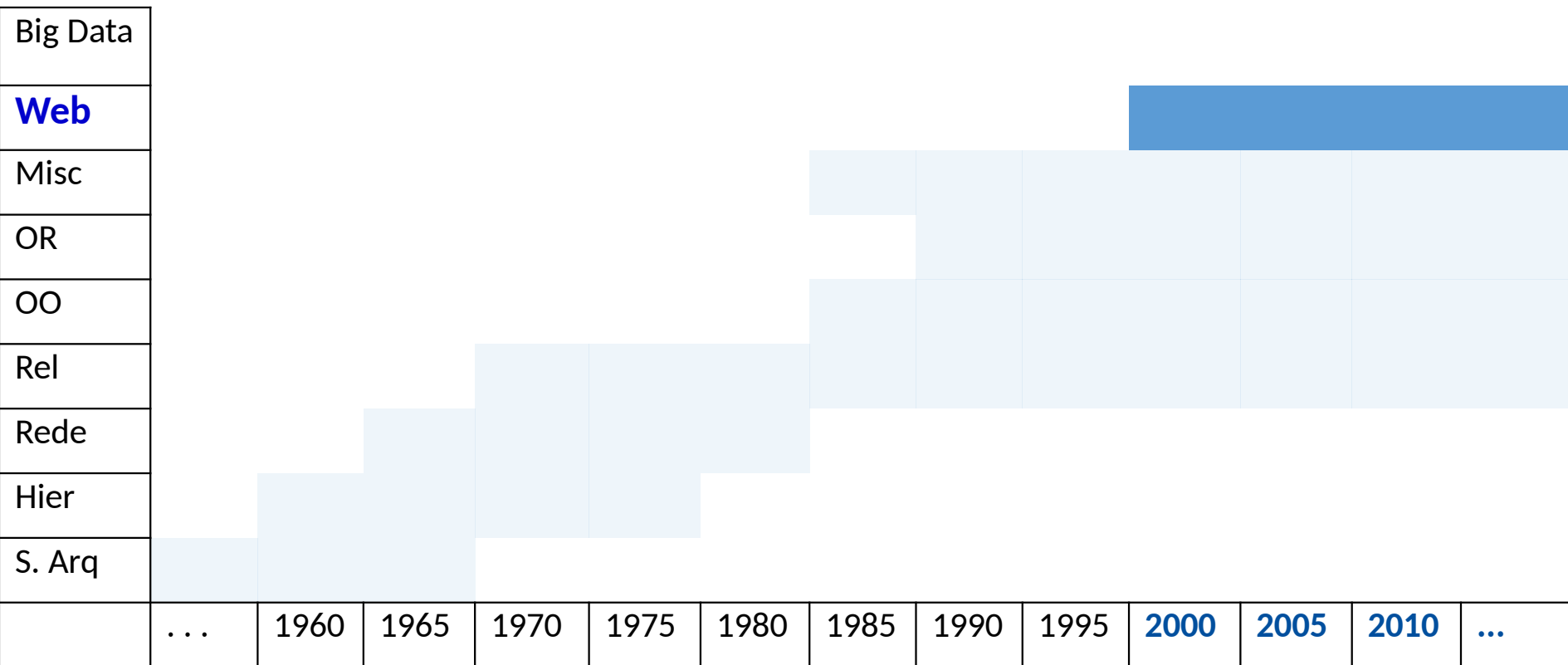
- Modelos de dados complexos (estrutura complexa + métodos)
- Invocação de métodos em consultas
- Herança de propriedades (atributos e métodos)

Evolução dos Modelos de BD



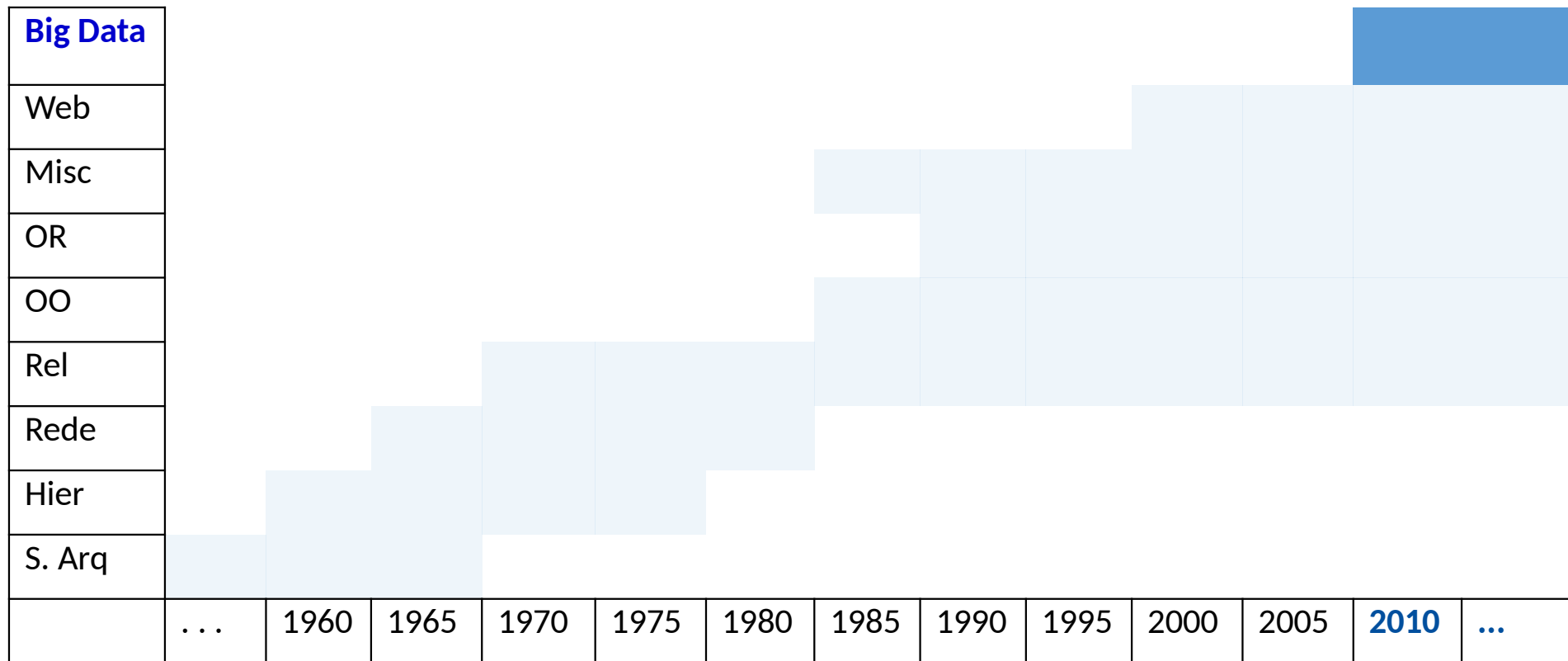
- Modelos de dados específicos para as necessidades de determinadas aplicações (dados geográficos, dados biológicos, dados multimídia, ...)

Timeline dos Modelos de BD



- Gerenciamento de dados com alta heterogeneidade (não-estruturados, semiestruturados e estruturados)
- Modelos de dados flexíveis

Timeline dos Modelos de BD



- Gerenciamento de dados heterogêneos e muito volumosos
- Modelos de dados mais simples
- Métodos de acesso limitados (minimizar *overhead* de gerenciamento; maximizar escalabilidade e disponibilidade)

Roteiro

1. Timeline dos Modelos de BD
2. Definição de Big Data
3. Os “X”s Vs da Big Data
4. Alguns Domínios de Aplicação
5. Considerações Finais

Definição de Big Data



1ª imagem recuperada pelo *Google Images* para a palavra-chave
“Big Data” (18/3/2024)

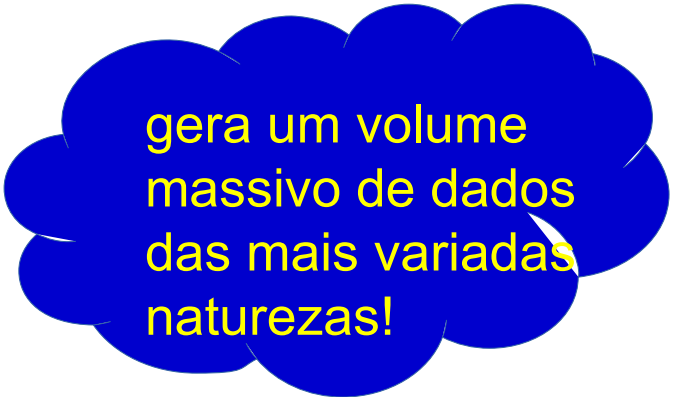
Por Quê “Big Data”?

- Modificação no uso e no tratamento da informação neste início de século XXI
 - Novo modelo **social**
 - Muita interação virtual
 - Novo modelo **econômico**
 - Muito comércio eletrônico
 - Novo modelo **tecnológico**
 - Muitas aplicações
(e aplicativos) Web



Por Quê “Big Data”?

- Modificação no uso e no tratamento da informação neste início de século XXI
 - Novo modelo **social**
 - Muita interação virtual
 - Novo modelo **econômico**
 - Muito comércio eletrônico
 - Novo modelo **tecnológico**
 - Muitas aplicações
(e aplicativos) Web



gera um volume
massivo de dados
das mais variadas
naturezas!

Big Data – Definições

- Falta de consenso (foco) para explicar o conceito...

(1) “Big Data é a quantidade enorme de informações nos servidores de bancos de dados”

(2) “Big Data requer um conjunto de técnicas e tecnologias com novas formas de integração de dados para revelar *insights* a partir de *data sets* que são diversos, complexos e em escala massiva”

(3) “Big Data são *data sets* tão grandes ou complexos que os *softwares* de processamento de dados tradicionais são inadequados para lidar com eles”

Big Data – “Meu Ponto de Vista”

Um *buzzword* que sinaliza um alerta para a comunidade de BD¹ (“*um movimento*”) no sentido de rever e aprimorar seus SGBDs e outras soluções associadas à gerência de dados, como aquelas voltadas à *BI* (*DWs, Data Mining, ...*), à descoberta e à integração de dados, visando atender novas demandas e desafios no tratamento de um universo cada vez maior de dados disponíveis em praticamente todos os domínios de aplicação

¹ Não apenas para a comunidade de BD, mas outras comunidades da Computação como IA, Computação Distribuída, Algoritmos e Complexidade, Redes, ...

Roteiro

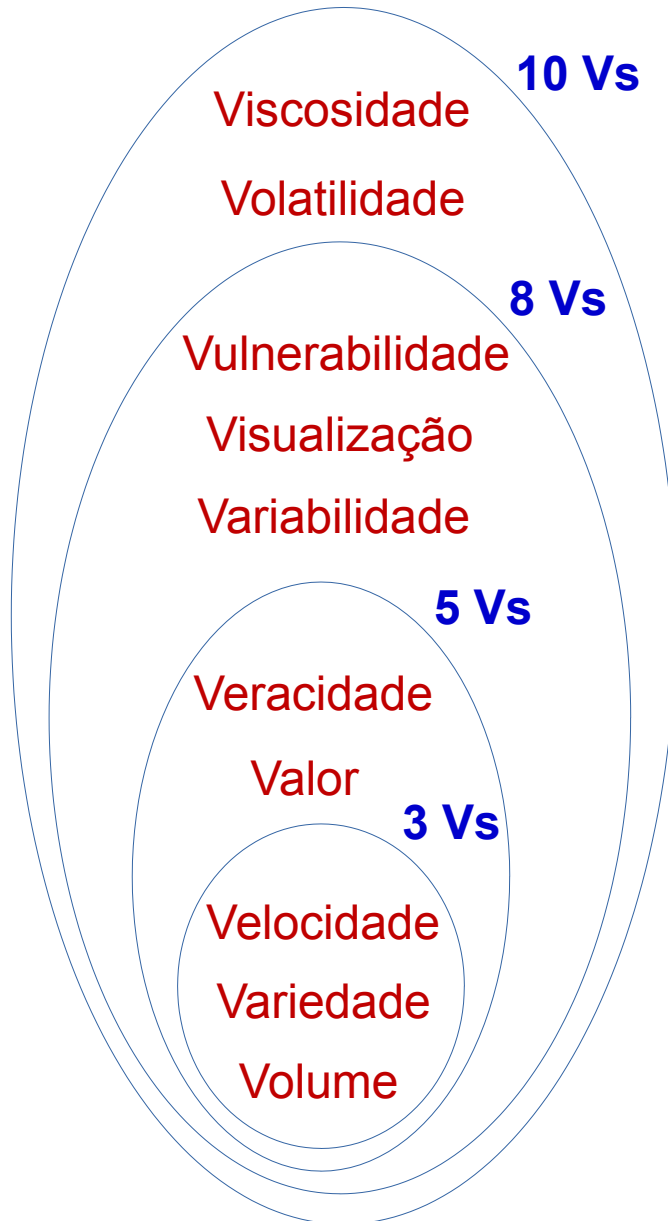
1. Timeline dos Modelos de BD
2. Definição de Big Data
3. Os “X”s Vs da Big Data
4. Alguns Domínios de Aplicação
5. Considerações Finais

“X”s Vs da Big Data

- Características (e ao mesmo tempo **desafios**) de Big Data
- Requisitos a serem considerados para um *framework* de gerenciamento de Big Data
- Evolução dos Vs (**não há consenso!**)
 - 3 Vs (requisitos fundamentais – *common framework* ou *core*)
 - 5 Vs
 - 8 Vs
 - 10 Vs

o quão relevante é cada um destes requisitos depende da intenção da aplicação...

“X”s Vs da Big Data



“X”s Vs da Big Data

10 Vs

Viscosidade

Volatilidade

- Desafio: lidar com a massiva quantidade de dados e informações que nos cercam hoje
- Objetivo: processar grandes volumes de dados para uma dada tarefa no menor tempo possível
- Desejável: soluções eficientes em termos de *storage* (p.ex.: *data centers*), com HW e SW robustos para garantir melhor processamento paralelo e escalabilidade para volumes de dados variados
- Exemplo: *Walmart* (> 2.5 Pb de dados de transações de usuários são coletados por hora)¹

Velocidade

Variedade

Volume

¹ Fonte: <https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/>

“X”s Vs da Big Data

10 Vs

Viscosidade

Vu

V

V

V

- Desafio: lidar com a natureza (logs de transações, textos do *Twitter*, vídeos de câmeras de monitoramento, ...) e heterogeneidade (dentro de uma mesma natureza) dos dados – **herança dos desafios do gerenciamento de dados na Web**
- Objetivo: processar dados necessários para uma dada tarefa independente da sua representação
- Desejável: soluções eficientes para *crawling*, extração, limpeza, determinação de similaridade e integração de dados **estruturados, semiestruturados e não-estruturados**
- Exemplo: *Twitter* (*posts* sobre determinado assunto podem ter uma infinidade de textos e *hashtags* possíveis)

Velocidad

Variedade

Volume

Dado Semiestruturado

- Dado com alguma estrutura explícita
 - parte não-estruturada composta por diferentes mídias (texto longo, imagem, ...)
 - exemplos:
 - páginas HTML de modo geral (incluindo resultados de pesquisas retornadas por navegadores Web)
 - documentos (e-mails, anúncios classificados, ...)

Dado Semiestruturado

CLASSIFICADOS

MORTGAGE SOLUTIONS!
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

3BR - Single Family \$22,900
Single Family 3 Bedroom 1 1/2 Bath Property has been completely renovated.

WANT TO OWN
If you cannot qualify for a traditional mortgage through the banks, we can help.

IMÓVEIS
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

ATTENTION HOME OWNER
2 Bedroom Property has been completely renovated and is currently occupied. Rent currently at \$600. Call us today at \$400.

EXCELLENT INVESTMENT
2 Family (Duplex) Units with 5 bedrooms and 3 full Baths. Property has been completely renovated and is currently occupied. Rent currently at \$1400. Call us today at \$400.

COMMERCIAL LOANS
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

NO MONEY DOWN
We help home buyers to purchase their home with ZERO money down. Call us today at \$400.

OFFICE AVAILABLE
3 individual enclosed office space with desks, at a great location, with parking and general facilities. Rental \$1000. Call us today at \$400.

LOOKING FOR INVESTORS
Looking for additional investors (experienced) and people located in Real Estate Market.

GENERAL HELP WANTED
Seeking someone for general help, eg. typing, organizing, etc. Call us today at \$400.

GARDEN MAINTENANCE
Experienced in maintaining lawns, shrubs, and trees. Call us today at \$400.

REPAIRS NEEDED
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

ATTENTION HOME OWNER
2 Bedroom Property has been completely renovated and is currently occupied. Rent currently at \$600. Call us today at \$400.

IMÓVEIS
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

ATTENTION HOME OWNER
2 Bedroom Property has been completely renovated and is currently occupied. Rent currently at \$600. Call us today at \$400.

EXCELLENT INVESTMENT
2 Family (Duplex) Units with 5 bedrooms and 3 full Baths. Property has been completely renovated and is currently occupied. Rent currently at \$1400. Call us today at \$400.

COMMERCIAL LOANS
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

NO MONEY DOWN
We help home buyers to purchase their home with ZERO money down. Call us today at \$400.

OFFICE AVAILABLE
3 individual enclosed office space with desks, at a great location, with parking and general facilities. Rental \$1000. Call us today at \$400.

LOOKING FOR INVESTORS
Looking for additional investors (experienced) and people located in Real Estate Market.

GENERAL HELP WANTED
Seeking someone for general help, eg. typing, organizing, etc. Call us today at \$400.

REPAIRS NEEDED
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

ATTENTION HOME OWNER
2 Bedroom Property has been completely renovated and is currently occupied. Rent currently at \$600. Call us today at \$400.

IMÓVEIS
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

ATTENTION HOME OWNER
2 Bedroom Property has been completely renovated and is currently occupied. Rent currently at \$600. Call us today at \$400.

EXCELLENT INVESTMENT
2 Family (Duplex) Units with 5 bedrooms and 3 full Baths. Property has been completely renovated and is currently occupied. Rent currently at \$1400. Call us today at \$400.

COMMERCIAL LOANS
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

NO MONEY DOWN
We help home buyers to purchase their home with ZERO money down. Call us today at \$400.

OFFICE AVAILABLE
3 individual enclosed office space with desks, at a great location, with parking and general facilities. Rental \$1000. Call us today at \$400.

LOOKING FOR INVESTORS
Looking for additional investors (experienced) and people located in Real Estate Market.

GENERAL HELP WANTED
Seeking someone for general help, eg. typing, organizing, etc. Call us today at \$400.

REPAIRS NEEDED
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

ATTENTION HOME OWNER
2 Bedroom Property has been completely renovated and is currently occupied. Rent currently at \$600. Call us today at \$400.

IMÓVEIS
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

ATTENTION HOME OWNER
2 Bedroom Property has been completely renovated and is currently occupied. Rent currently at \$600. Call us today at \$400.

EXCELLENT INVESTMENT
2 Family (Duplex) Units with 5 bedrooms and 3 full Baths. Property has been completely renovated and is currently occupied. Rent currently at \$1400. Call us today at \$400.

COMMERCIAL LOANS
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

NO MONEY DOWN
We help home buyers to purchase their home with ZERO money down. Call us today at \$400.

OFFICE AVAILABLE
3 individual enclosed office space with desks, at a great location, with parking and general facilities. Rental \$1000. Call us today at \$400.

LOOKING FOR INVESTORS
Looking for additional investors (experienced) and people located in Real Estate Market.

GENERAL HELP WANTED
Seeking someone for general help, eg. typing, organizing, etc. Call us today at \$400.

REPAIRS NEEDED
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

ATTENTION HOME OWNER
2 Bedroom Property has been completely renovated and is currently occupied. Rent currently at \$600. Call us today at \$400.

IMÓVEIS
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

ATTENTION HOME OWNER
2 Bedroom Property has been completely renovated and is currently occupied. Rent currently at \$600. Call us today at \$400.

EXCELLENT INVESTMENT
2 Family (Duplex) Units with 5 bedrooms and 3 full Baths. Property has been completely renovated and is currently occupied. Rent currently at \$1400. Call us today at \$400.

COMMERCIAL LOANS
We provide complete Property Management, Leasing and Real Estate Services for Residential and Commercial Properties.

NO MONEY DOWN
We help home buyers to purchase their home with ZERO money down. Call us today at \$400.

OFFICE AVAILABLE
3 individual enclosed office space with desks, at a great location, with parking and general facilities. Rental \$1000. Call us today at \$400.

LOOKING FOR INVESTORS
Looking for additional investors (experienced) and people located in Real Estate Market.

GENERAL HELP WANTED
Seeking someone for general help, eg. typing, organizing, etc. Call us today at \$400.

<anuncio>
<transacao>Vendo</transacao>, por motivo de viagem, <produto>automóvel Gol I 97</produto>, cor azul, em ótimo estado de conservação. Preço: R\$<preco>9000,00</preco>. Tratar com <contato><nome>Pedro</nome> fone</fone> 99991111</fone></contato>
</anuncio>
<anuncio>
Atenção! Se você deseja vender o seu veículo, nós realizamos o melhor negócio. <transacao>Compramos</transacao> qq tipo de <produto>veículo</produto>. Ligue-nos: <contato><fone>32340011</fone> ou envie um e-mail:<eMail>lojao@bla.com.br</eMail><contato></anuncio>

“X”s Vs da Big Data

10 Vs

- Desafio: lidar com o consumo/geração de Big Data em alta velocidade - alta taxa de fluxo de dados no sistema (**fast data**)
- Objetivo: processar dados necessários para uma dada tarefa independente da sua taxa de recebimento
- Desejável: soluções para a melhoria de canais de transmissão (redes de fibra ótica, uso de satélites, emissores de sinais de alta capacidade) e soluções inteligentes para processamento em tempo real
- Exemplo: *Walmart* (> 1 milhão de transações de clientes / hora)¹

Veracidade

Variedade

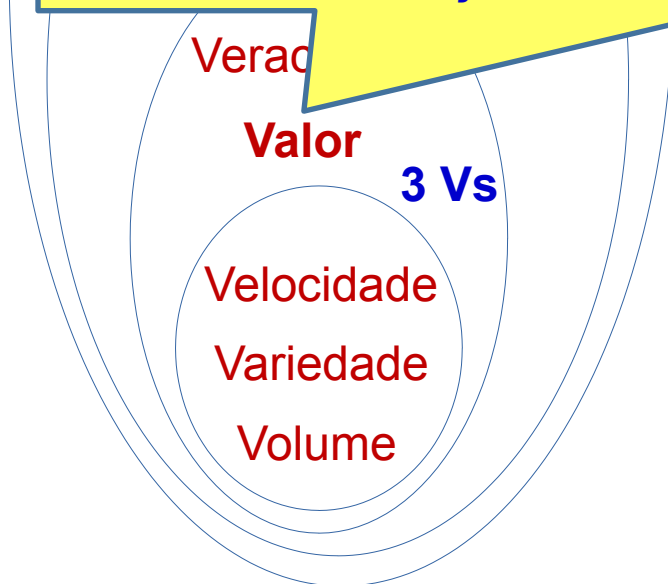
Vs

Velocidade

Variedade

Volume

- Desafio: considerar o valor agregado ao dado (benefício para a sociedade), que geralmente diz respeito ao: (i) seu uso analítico; (ii) sua habilidade de ser útil na geração de novos produtos/serviços ou nos seus aprimoramentos (**Big Data Analytics**)
- Objetivo: realizar operações analíticas eficientes sobre dados com potencial para relevar informação relevante para a tomada de decisões
- Desejável: (i) uso analítico: descoberta de *insights* relevantes escondidos em dados custosos de processar, utilizando, por exemplo, técnicas de Mineração de Dados e Aprendizado de Máquina; (ii) geração de novos produtos/serviços: habilidade de correlação de dados para oferecer soluções com melhor qualidade
- Exemplo: previsão de desastres naturais em uma região com base na análise e correlação de dados climáticos, sísmicos, jornalísticos, ...



- Desafio: considerar a qualidade dos dados
- Objetivo: avaliar o grau de confiança (ou de incerteza) de um conjunto de dados e eliminar dados com baixo grau de confiança (***data cleaning***)
- Desejável: desenvolver técnicas para verificar se amostras de dados fazem sentido (mantém um padrão de coerência em termos de conteúdo? estão completos em sua grande maioria?), a reputação da procedência dos dados (existe verificação de integridade dos dados gerados naquela fonte de dados? As regras de integridade estão corretas?), ...
- Exemplos: (i) detecção de avaliações mal intencionadas (*fake reviews*) em sites de avaliação de produtos por apresentarem comentários que não fazem sentido, por estarem muito fora do padrão; (ii) *Google Flu Trends* estimou 2x mais casos de *influenza* do que o reportado oficialmente pelo *CDC (Centers for Disease Control and Prevention)*² – descontinuado em 2015... :-)

Veracidade

Valor

3 Vs

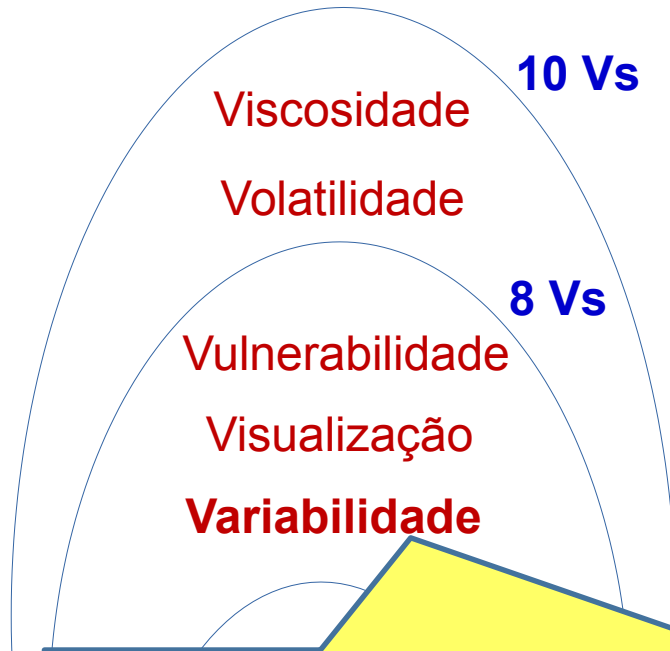
Velocidade

Variedade

Volume

² Fonte: <https://www.nature.com/articles/494155a>

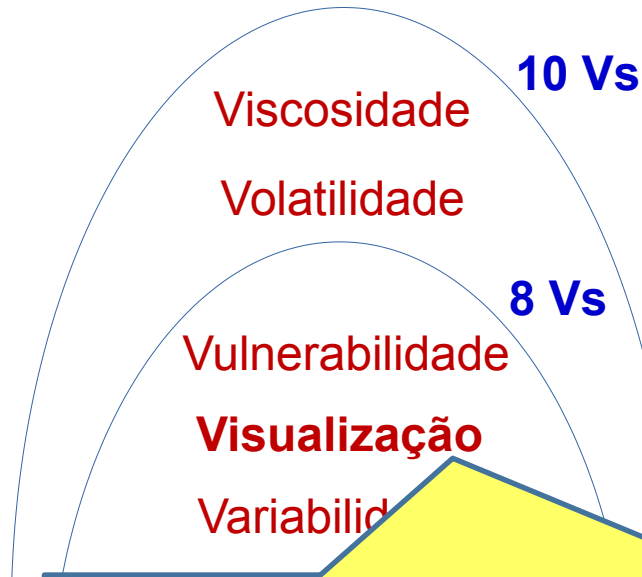
“X”s Vs da Big Data



- Desafio: lidar com variações nos 3 Vs, ou seja, picos de alto e baixo volume, variedade e velocidade (situações não-determinísticas)
- Objetivo: garantir que o desempenho no processamento de Big Data não seja comprometido com tais variações
- Desejável: desenvolver soluções que garantam **elasticidade** no tratamento de Big Data, como a adoção de serviços nas nuvens
- Exemplo: *Amazon elastic compute cloud* (serviços em diferentes níveis: (de infraestrutura a gerenciadores de dados com modelos flexíveis)³

³ <https://aws.amazon.com/pt/ec2/>

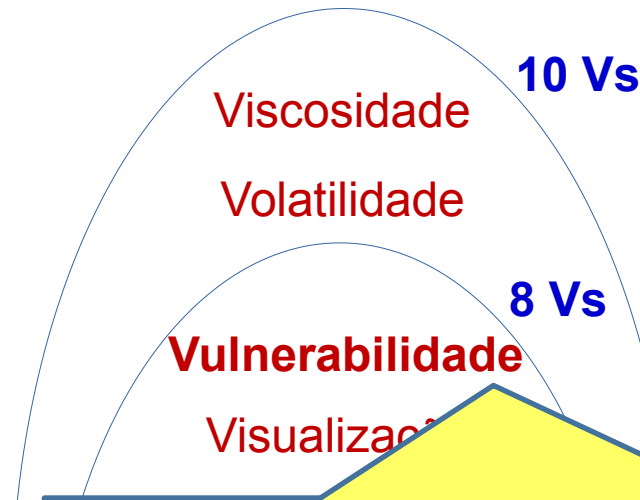
“X”s Vs da Big Data



- Desafio: lidar com a complexidade de visualizar Big Data volumosos, variados e processados em alta velocidade
- Objetivo: garantir que os dados sejam visualizados e bem compreendidos pelos usuários e tomadores de decisão
- Desejável: desenvolver técnicas de **visualização científica** adequados ao seu Big Data (gráficos, grafos, *browsers*, ...)
- Exemplo: *Logi Analytics* (aplicação para *Big Data visual analytics*)⁴

⁴ <https://insightsoftware.com/logi-analytics/>

“X”s Vs da Big Data



- Desafio: manter Big Data livre de ataques e falhas durante a sua manipulação
- Objetivo: garantir Big Data sempre seguro
- Desejável: desenvolver técnicas de segurança eficientes para dados
- Exemplos: técnicas de *recovery* e criptografia adaptadas à Big Data

“in May 2016, a hacker called Peace posted data on the dark web to sell, which allegedly included information on 167 million LinkedIn accounts and ... 360 million emails and passwords for MySpace users.” (LinkedIn Vulnerability)

Roteiro

1. Timeline dos Modelos de BD
2. Definição de Big Data
3. Os “X”s Vs da Big Data
4. Alguns Domínios de Aplicação
5. Considerações Finais

Social Networks

- Dados de várias naturezas
 - *Posts* (textos), imagens, vídeos, georeferenciados, ...
- Complexa rede de relacionamentos
 - amizade, grupos, eventos, ...
- Muitos acessos e inserções



Internet of Things (IoT)

- Objetivo: integração do mundo físico com sistemas computacionais visando melhor eficiência e precisão em inúmeras tarefas com intervenção humana reduzida
- Aplicações:
 - *smart homes* (controle de tarefas domésticas – Exemplos: iluminação, ar condicionado, segurança, eletrodomésticos)
 - *smart cities* (monitoramento e controle de dispositivos em locais públicos – Exemplos: irrigação de jardins, iluminação pública, semáforos)



Internet of Things (IoT)

- Principais desafios
 - Monitoramento, análise e controle de múltiplos sensores que geram dados de natureza diversa (*realtime streaming data analytics*)
 - Quanto maior a área urbana maior o seu Big Data...
 - Dados devem ser precisos e confiáveis!
- Estima-se que 50 bilhões de dispositivos estarão conectados à Internet no início da 2ª década do século XXI

e-Commerce

- Grandes acervos de produtos
 - **Dados multimídia** (fotos, texto descritivo, metadados, avaliações, ...)
- Grande volume de **transações**
- **Análise de vendas e recomendação** de novos produtos
 - Considera perfil do usuário e de usuários similares, similaridade de produtos, avaliações dos produtos, ...



e-Commerce



- Exemplo: [Amazon.com](https://www.amazon.com)
 - Comércio eletrônico
 - Serviços para Computação nas Nuvens
 - Alguns números
 - Mais de 650 milhões de visitas ao seu *Website* por ano
 - > 130 milhões de consumidores por mês
 - Consumidores de mais de 170 países
 - Transações de vendas em período de Natal: > 1 bilhão de itens

Healthcare

- Principais Objetivos
 - Análise preventiva do quadro clínico das pessoas visando evitar problemas de saúde
 - Busca de cura para doenças
 - Predição de epidemias
- Problemática cada vez mais relevante
 - Crescimento da população mundial
 - Pessoas estão vivendo mais



Big Data in Healthcare

- Suporte a **Sistemas de Apoio a Diagnósticos**
 - Médicos podem realizar análises complexas com base no **cruzamento de dados** de pacientes provenientes de **múltiplas fontes**, em **grande volume** e com **múltiplos formatos**
 - Cadastros em BDs convencionais
 - Sensores de monitoramento (muitas vezes contínuo) do quadro clínico
 - Fixos (dispositivos conectados a pacientes internados)
 - Móveis (dispositivos *fitness*, medidores de glicose, calorias, ...)
 - Imagens (tomografias, ...) ⁵<https://www.acor.org/>
 - Redes sociais para *Healthcare* (exemplo: ACOR⁵ – rede nos EUA com > 100 mil pacientes organizados em grupos de prevenção do câncer)
 - Aplicativos que analisam áudios com falas de pacientes e sugerem sintomas como depressão e derrame ⁶<https://enterprises.upmc.com/phda/>
 - Cruzamento com dados de pacientes com sintomas similares, incluindo dados genéticos (exemplo: iniciativa *Pittsburgh Health Data Alliance*⁶)

Roteiro

1. Timeline dos Modelos de BD
2. Definição de Big Data
3. Os “X”s Vs da Big Data
4. Alguns Domínios de Aplicação
5. Considerações Finais

Considerações Finais

- Big Data – o que é ?
 - *Buzzword* que remete a um *revival* dos grandes problemas de gerenciamento de dados (modelagem, acesso eficiente, integração, indexação, similaridade, ...), que devem ser *revistos* para lidar com a magnitude dos x's Vs
 - Desafios: captura, armazenamento, análise, organização e integração, compartilhamento, visualização, consulta, atualização e privacidade dos dados
 - Uma nova dinâmica para fluxos informacionais para interação entre sociedade, governos e serviços em geral

Considerações Finais

- Big Data – benefícios
 - Seu gerenciamento garante maior disponibilidade de dados & informações úteis para consumo humano
 - O interesse neste assunto tem impulsionado a pesquisa & desenvolvimento de novas soluções
 - Computação nas nuvens
 - Tecnologias para otimização de seu processamento, como Hadoop, HDFS, BDs NoSQL e outros tipos, *Analytics*, ...

UFSC-CTC-INE-PPGCC

INE 410131 – Gerencia de Dados para Big Data

Aula 2 – Introdução à Big Data

Ronaldo S. Mello

2024/1