



Apresentação da Disciplina e Conceitos Iniciais

Curso: Tecnologia em Gestão da Tecnologia da Informação

Prof.: Jean Henrique de Sousa Câmara

Contato: jean.camara@ifsudestemg.edu.br

Big Data

Apresentação da disciplina

2

NOME DA DISCIPLINA: BIG DATA

Período: Condicionada à oferta do docente

Carga Horária: 66 horas

Natureza: optativa

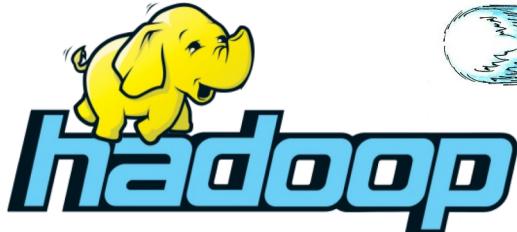
Ementa:

Definição e fundamentos do Big Data. Utilização do Big Data. Tecnologia para o Big Data. Capacitação e profissionais. Estágios da preparação de dados: Acessando; Auditando; Melhorando e enriquecendo os dados; Determinando a estrutura de dados; Pesquisando os dados; e Modelando. Combinando dados de múltiplas fontes. Confidencialidade e ética ao manipular dados. Manipulando dados desalinhados, inconsistentes e não padronizados. Transformação e transferência de dados. Ferramentas para preparação de dados.

Quais as suas experiências?

3

- Está gostando do curso?
 - Qual o conteúdo que mais gostou até o momento?
- Quais dessas ferramentas/tecnologias você conhece?



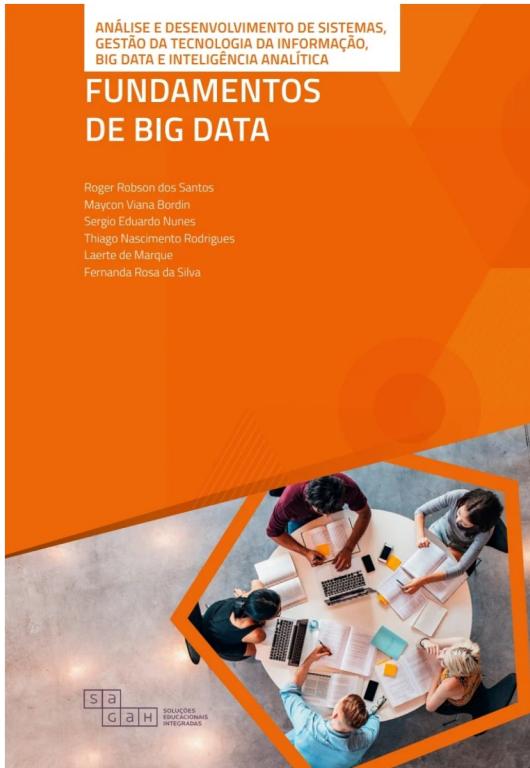
Avaliações

4

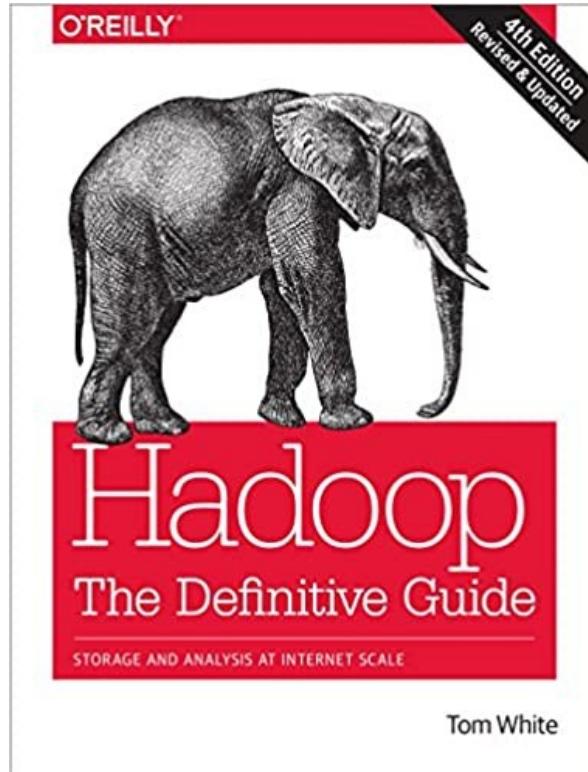
- Avaliação I (2 pontos)
- Avaliação II (2 pontos)
- Atividades em sala de aula (2 pontos)
- Trabalhos (4 pontos)

Referências bibliográficas

5



SANTOS, R. R. et al.
Fundamentos de Big Data.
SAGAH, Porto Alegre - RS, 2021.



WHITE, T. **Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale.** Editora O'Reilly Media, 4.ed. Paris, 2015.



VIDA, E. S. et al. **Data warehouse.** SAGAH, Porto Alegre - RS, 2021.

Introdução

6

- Big Data é o termo usado para se referir a **enormes conjuntos de dados** coletados de **várias fontes**
- Não seria possível processar esses conjuntos de dados manualmente e nem por banco de dados ou aplicações de processamento tradicionais

Cada ser humano criou cerca de 1.7 MB de dados por segundo em 2020

Introdução

7

➤ Unidade de armazenamento

Bit (b*)	1 unidade
Byte (B*)	8 bits
Kilobyte (KB)	1024 Byte
Megabyte (MB)	1024KB
Gigabyte (GB)	1024 MB
Terabyte (TB)	1024 GB
Petabyte (PB)	1024 TB
Exabyte (EB)	1024 PB
Zettabyte (ZB)	1024 EB
Yotabyte (YB)	1024 ZB

A previsão é de que ainda neste ano sejam gerados 35 Zettabytes de dados

Tipos de dados em um SGBD

8

- Tipos de dados em um sistema de gerenciamento de banco de dados (SGBD)
 - Estruturados
 - Não estruturados
 - Semiestruturados

Levaria 181 milhões de anos para baixar todos os dados da Internet

Tipos de dados em um SGBD

9

➤ Dados estruturados

- São aqueles com tamanhos definidos em seu desenvolvimento
- Em grande parte, correspondem a números, datas e palavras
- Geralmente são utilizados em bancos de dados do tipo relacional

```
4 ▼ CREATE TABLE Meus_Contatos (
5     Id int PRIMARY KEY auto_increment,
6     Nome varchar(40),
7     DataNasc date NOT NULL,
8 );
```

Tipos de dados em um SGBD

10

➤ **Dados não estruturados**

- São dados que não possuem formatos e cujo tamanho pode variar
- São encontrados em fotografias, vídeos, imagens de satélites, mídias sociais, entre outros



Tipos de dados em um SGBD

11

➤ Dados semiestruturados

- São considerados como um meio termo entre os dados estruturados e os dados não estruturados
- O seu uso está ligado a aplicações web, em que os dados são convertidos em tags

```
1 ▼ <?xml version="1.0" encoding="ISO-8859-1"?>
2 ▼   <recado>
3     <para>Aluno</para>
4     <de>Professor Serginho</de>
5     <título>Vamos estudar?</título>
6     <corpo>Fundamentos de Big Data</corpo>
7   </recado>
```

Os 5 Vs do Big Data

12



Os 5 Vs do Big Data

13

1. Volume

- Nesse exato momento, uma verdadeira enxurrada de dados é gerada para nortear indivíduos, empresas e governos
- Esses dados são advindos de diversas fontes, como redes sociais, motores de busca da internet, e-commerce...
- Qualquer objeto conectado à Internet pode gerar dados (Internet das Coisas IoT)
- A produção de dados dobra a cada dois anos

Existem cerca de 6 bilhões de telefones móveis no planeta

Os 5 Vs do Big Data

14

2. Velocidade

- Se refere à velocidade de **captação, organização e análise** de dados
- Os dados devem ser processados rapidamente para que possam ser usados em tempo hábil
 - Pode ser mais interessante ter uma quantidade um pouco menor de dados em tempo real, do que uma enorme quantidade que só poderá ser disponibilizada para uso depois de um tempo considerável

30 bilhões de imagens são publicadas por mês no Instagram

Os 5 Vs do Big Data

15

3. Variedade

- Os dados provêm de diversas fontes: redes sociais, aplicativos, e-mails, GPS, IoT, bancos de dados públicos, revendedores autorizados...
 - Não seguem os mesmos padrões e nem fornecem os mesmos tipos de informação
- Aproximadamente 80% dos dados são não estruturados ou estão em diferentes formatos, o que dificulta a análise
- A variedade de fontes de dados só tende a aumentar com o avanço tecnológico

Os 5 Vs do Big Data

16

4. Veracidade

- Esse talvez seja um desafio encontrado no Big Data
 - Com tantos dados disponíveis, como separar os verdadeiros dos falsos?
 - Faz-se necessário verificar as fontes, os dados tendenciosos e as datas de publicação

Estima-se que 3.1 trilhões de dólares sejam desperdiçados por ano devido a problemas de qualidade dos dados

Os 5 Vs do Big Data

17

5. Valor

- Transformar um verdadeiro tsunami de informações em dados que efetivamente podem ser utilizados nos negócios
- Os dados devem trazer **informações relevantes**, que possam proporcionar um diferencial de mercado ou, ainda, auxiliar os gestores na tomada de decisão

Big Data x Ciência dos Dados

18

- Big Data e Ciência dos Dados são a mesma coisa?



Big Data x Ciência dos Dados

19

- Big Data e Ciência dos Dados são a mesma coisa?
 - Não
- **Big Data** é a matéria-prima, ou seja, dados
- **Ciência dos Dados** é um conjunto de técnicas para análise de dados
- Quando aplicamos Ciência de Dados ao Big Data extraímos valor e então temos o que é chamado de ***Big Data Analytics***

Etapas do Big Data Analytics

20

1. Coleta de Dados
2. Integração dos Dados
3. Análise e Modelagem dos Dados
4. Interpretação das Informações

Coleta de Dados

21

- Também chamada de aquisição ou gravação de dados
- É a fase de **reunir** todo aquele grande volume e diversidade de **dados**
- É necessário que esses dados já passem por algum tipo de **filtragem ou formatação**, eliminando erros e dados duplicados e incompletos

Coleta de Dados

22

- No caso do marketing para o varejo, por exemplo, alguns dados podem ser coletados no e-commerce da empresa, tais como:
 - cliques em anúncios
 - tipo de dispositivo (smartphone, notebook etc.)
 - sistema operacional do dispositivo
 - login do cliente no site, se houver
 - endereço de IP
 - endereço de e-mail
 - localização
 - histórico de buscas
 - histórico de compras do cliente

Integração dos Dados

23

- Como os dados são de fontes, formatos e características diferentes, devem receber **tratamentos específicos**
- Devem ser definidos critérios de **validação, aceitação, segurança e categorias dos dados**, conforme as suas fontes
- No varejo, por exemplo, análises de comportamento do consumidor do e-commerce podem ser combinadas com outras informações de lojas físicas para gerar insights

Análise e Modelagem dos Dados

24

- Uma das fases mais importantes
 - onde os dados começam a **ganhar valor** e se transformar em informação
- É necessário ter profissionais capacitados e o suporte de tecnologias de **inteligência artificial e machine learning**
- É importante também pesquisar novos tipos de **visualização de dados** para que sejam feitas descobertas valiosas

Interpretação das Informações

25

- É a última fase e também aquela que faz valer todo o investimento em big data
- Serão extraídos *insights* que vão garantir ao negócio **diferenciais competitivos** e oferecer uma ótima experiência ao cliente

Desafios do Big Data

Desafios do Big Data

27

- **Infraestrutura**
 - Armazenamento
 - É necessário que exista uma infraestrutura de tecnologia da informação (TI) adequada ao cenário que se deseja explorar
 - Muitas empresas estão migrando para os serviços em nuvem
 - Infraestrutura de redes de computadores
 - Há uma limitação quanto à taxa de processamento e transmissão de dados

Desafios do Big Data

28

- **Gerenciamento do crescimento dos dados**
 - Estima-se que, diariamente, as empresas produzam por volta de 2.5 EB de dados
 - Os dispositivos (IoT) produzem cerca de 5 EB de dados por dia
 - Aproximadamente 100 sensores estão instalados nos carros modernos
 - Para captar e armazenar esses dados é necessário usar algumas técnicas
 - Compactação
 - Deduplicação
 - Hierarquização

Desafios do Big Data

29

- **Extração e tratamento dos dados**
 - Identificar quais dados são interessantes e essenciais para a empresa
 - Diferentes formatos
- **Proteção e segurança**
 - Proteção dos dados da empresa: devido aos métodos criptográficos, os dados capturados não podem ser utilizados para análise de dados
 - Proteção das informações do big data: dados já filtrados requerem proteção, pois refletem um precioso bem para a empresa

Desafios do Big Data

30

➤ Resistência organizacional

- Além dos desafios tecnológicos, existem aqueles ligados aos problemas administrativos encontrados em grande parte das organizações
 - Falta de alinhamento organizacional
 - Falta de entendimento
 - Resistência gerencial

Desafios do Big Data

31

➤ Profissionais capacitados

- Há a necessidade de profissionais que saibam operacionalizar, desenvolver e configurar os software e toda a infraestrutura física para mineração e tratamento dos dados
- Além de possuírem competências e habilidades no desenvolvimento de tecnologias é necessário que os profissionais **saibam estreitar seu relacionamento com os gestores**, permitindo que os dados gerem *insights* confiáveis e proporcionem uma tomada de decisão estratégica

Aplicação do Big Data

Aplicação do Big Data

33

- Em quais segmentos o big data pode ser aplicado?
- Quais são os *insights* esperados pelos gestores?
- Conhecer exemplos de algumas aplicações dessa tecnologia pode proporcionar ao profissional de TI maior assertividade no momento de aplicar as técnicas do big data em diversos cenários

Aplicação do Big Data

34

- **Ramo empresarial**
 - Comportamentos e tendências
 - quais produtos ou serviços podem ser direcionados a determinado nicho ou grupo (sistema de recomendação)
 - Estratégia de marketing
 - realizar análises de dados e direcionar as ofertas de produtos e serviços de forma mais assertiva (*remarketing*)
 - Melhoria de produtos e serviços
 - *feedbacks* fornecidos por consumidores em redes sociais, sites de reclamação...

Aplicação do Big Data

35

- **Área da saúde**
 - Medicina de precisão
 - corrigir as deficiências na prescrição de medicamentos que funcionam bem para determinados pacientes e para outros não
 - Prontuários eletrônicos
 - prescrição de medicamentos, diagnósticos baseados em históricos similares, acesso a documentos em diferentes centros de saúde

Aplicação do Big Data

36

- **Trânsito em grandes cidades**
 - O big data é capaz de gerar informações de possíveis pontos de congestionamento

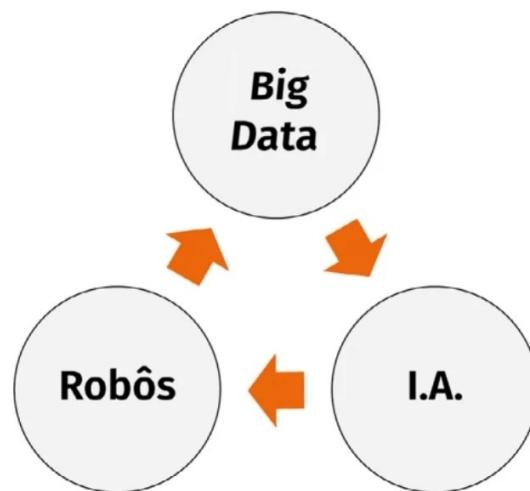


Aplicação do Big Data

37

➤ **Investimentos**

- Esse cenário talvez represente uma das maiores aplicabilidades do big data



1 TB de informação é criada durante uma única sessão da bolsa de valores americana (NYSE)

Empresas que trabalham com Big Data

38



[facebook](#) [Instagram](#) [WhatsApp](#) [Messenger](#) [oculus](#)



[americanas.com](#)



mercado
pago



Vivo Ads

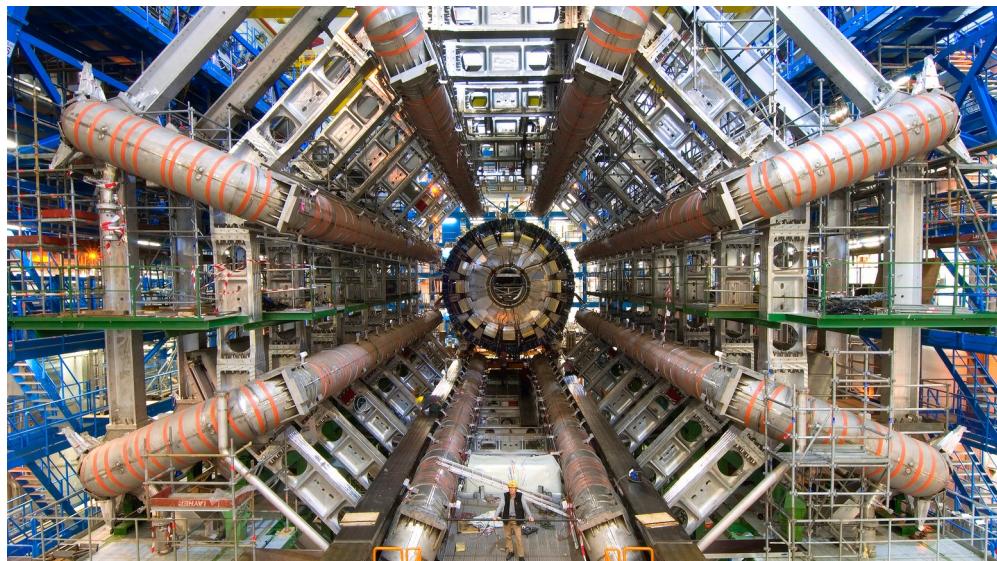
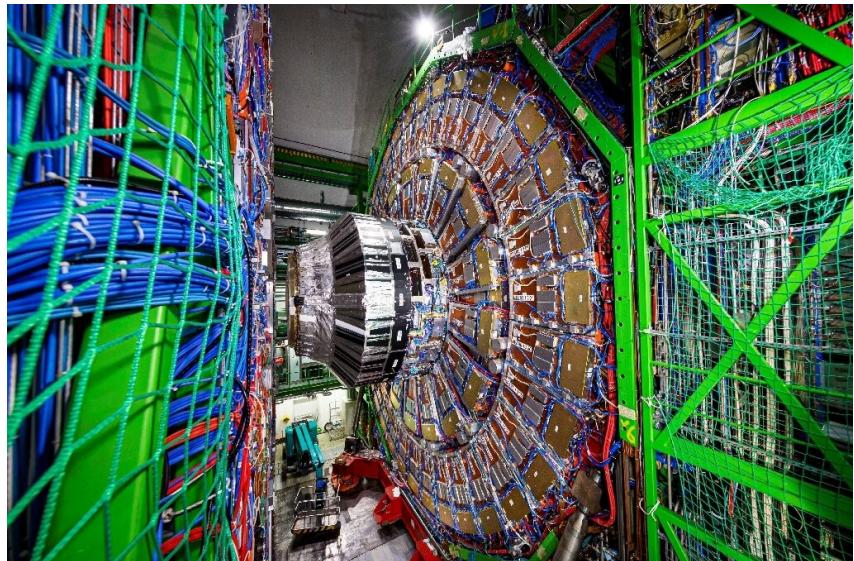
39

- Produto que fornece dados demográficos sobre as pessoas que passam por um determinado local e horário, à escolha do cliente
- É possível saber quantas pessoas transitam na área, a proporção de homens e mulheres, a distribuição por idade...
- 73M clientes: 43M deram consentimento para a empresa usar seu cadastro e 26M para o uso da localização

Projetos que produzem muitos dados

40

- The Large Hadron Collider (CERN)
 - Maior acelerador de partículas e o de maior energia no mundo
 - Cerca de 15 PB por ano



Projetos que produzem muitos dados

41

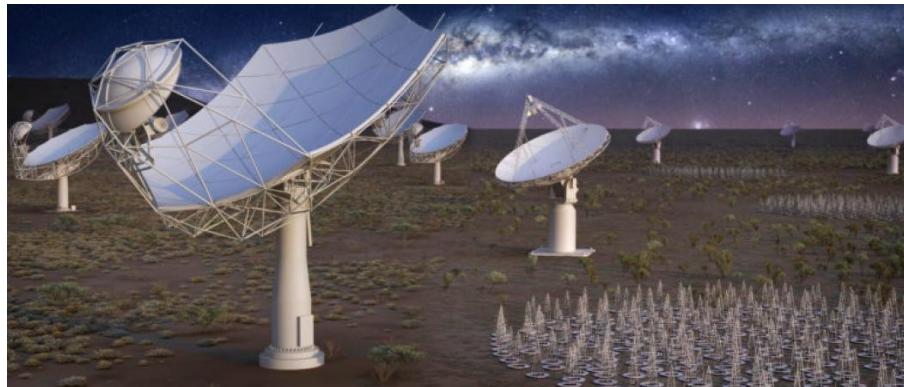
- Large Synoptic Survey Telescope (LSST)
 - Telescópio refletor com espelho de 8,4 metros
 - Vai fotografar o céu inteiro em algumas noites
 - 6-10 PB por ano



Projetos que produzem muitos dados

42

- Square Kilometre Array (SKA)
 - Será o maior telescópio do mundo, capaz de captar ondas de rádio
 - Quando estiver terminado, vai estar espalhado pela Austrália, Nova Zelândia, África do Sul e outros países africanos
 - 0,3 a 1,5 EB por ano



Aplicação do Big Data

43



- Não existe uma limitação para determinado segmento
- Desde que seja possível, de alguma forma, **gerar e captar dados**, será possível fazer com que o big data se torne uma importante ferramenta de estratégia de mercado

Mais estatísticas

44

- Estatísticas do Instagram 2022
- Lista de estatísticas de Big Data para 2022
- Estatísticas do Google 2022

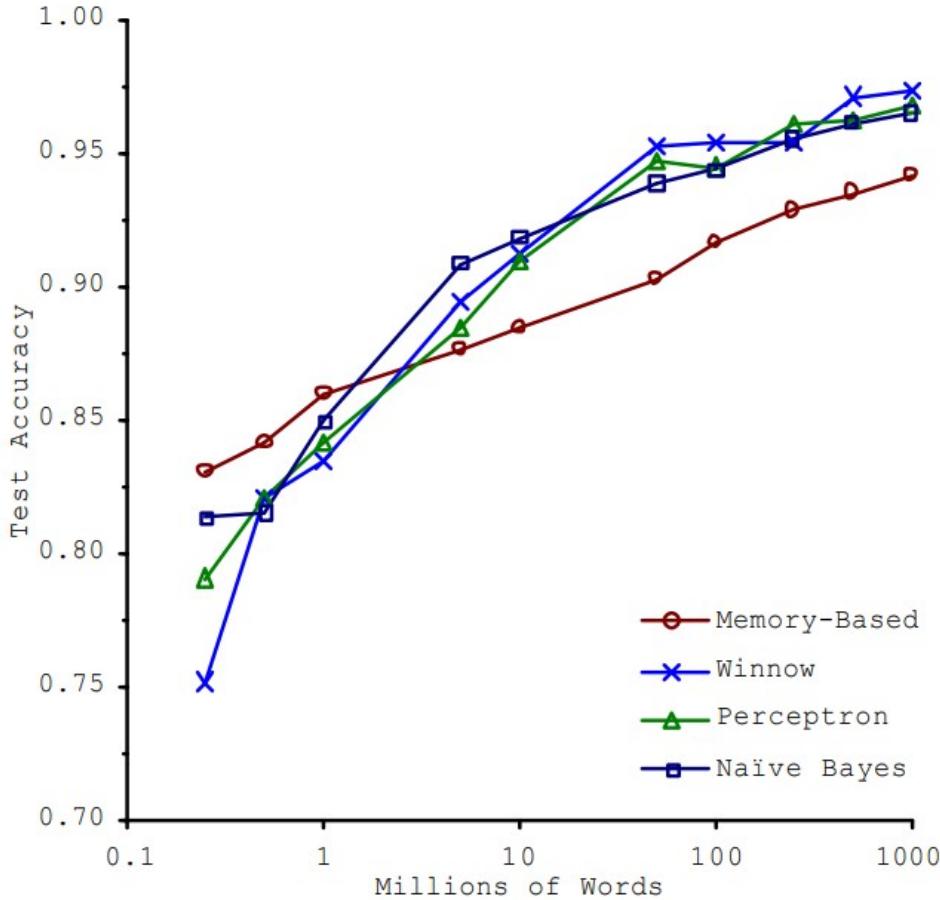
Quantidade de Dados em ML

45

- A quantidade de dados em Machine Learning faz diferença?
 - Sim, no treinamento mais dados fazem diferença!
 - The Unreasonable Effectiveness of Data
 - “Invariavelmente, modelos simples que usam grandes volumes de dados, superam modelos mais elaborados que utilizam menores volumes de dados” (HALEVY, NORVIG & PEREIRA, 2009)
 - Ex: na desambiguação de nomes, se A e B raramente ocorrem juntos, mas ambos geralmente ocorrem com C, então A e B podem ser sinônimos
 - Evidência significativa se observado em milhões de sentenças

Quantidade de Dados em ML

46



(BANKO & BRILL, 2001)

Armazenamento x Processamento

47

- Conjuntos de dados de TB são comuns e volumes na ordem de PB começam a surgir com muita frequência
- Tendência clara: Nossa capacidade de **gerar dados** está rapidamente **superando** nossa habilidade de **processar os dados** que armazenamos
- Mais preocupante: o aumento na capacidade de armazenamento está superando as melhorias em largura de banda
- **Está difícil até mesmo ler os dados** que estão sendo armazenados

Armazenamento x Processamento

48

- Capacidade dos discos passou de dezenas de MB na metade dos anos 80 para alguns TB hoje em dia
- Por outro lado, a latência e a largura de banda melhoraram relativamente pouco
- Latência melhorou 2x nos últimos 25 anos
- Banda talvez 50x

Tratamento de Big Data

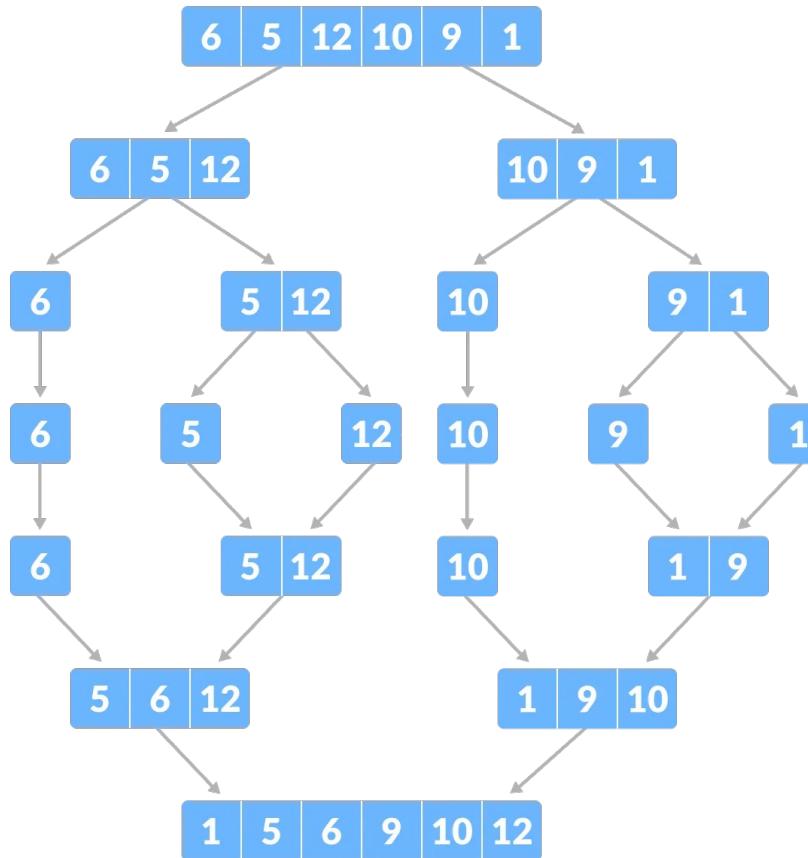
49

- Como que faremos para processar uma grande quantidade de dados se os dispositivos que temos não permitem tal tarefa?
- Usaremos o fundamento de **Divisão e Conquista**
 - **Divisão:** Se o tamanho da entrada é muito grande para a aplicação de uma solução simples, dividir o problema em dois ou mais subproblemas disjuntos
 - **Conquista:** Resolver o subproblema de tamanho adequado
 - **Combinação:** Tomar as soluções dos subproblemas e combiná-las para compor uma solução para o problema original

Tratamento de Big Data

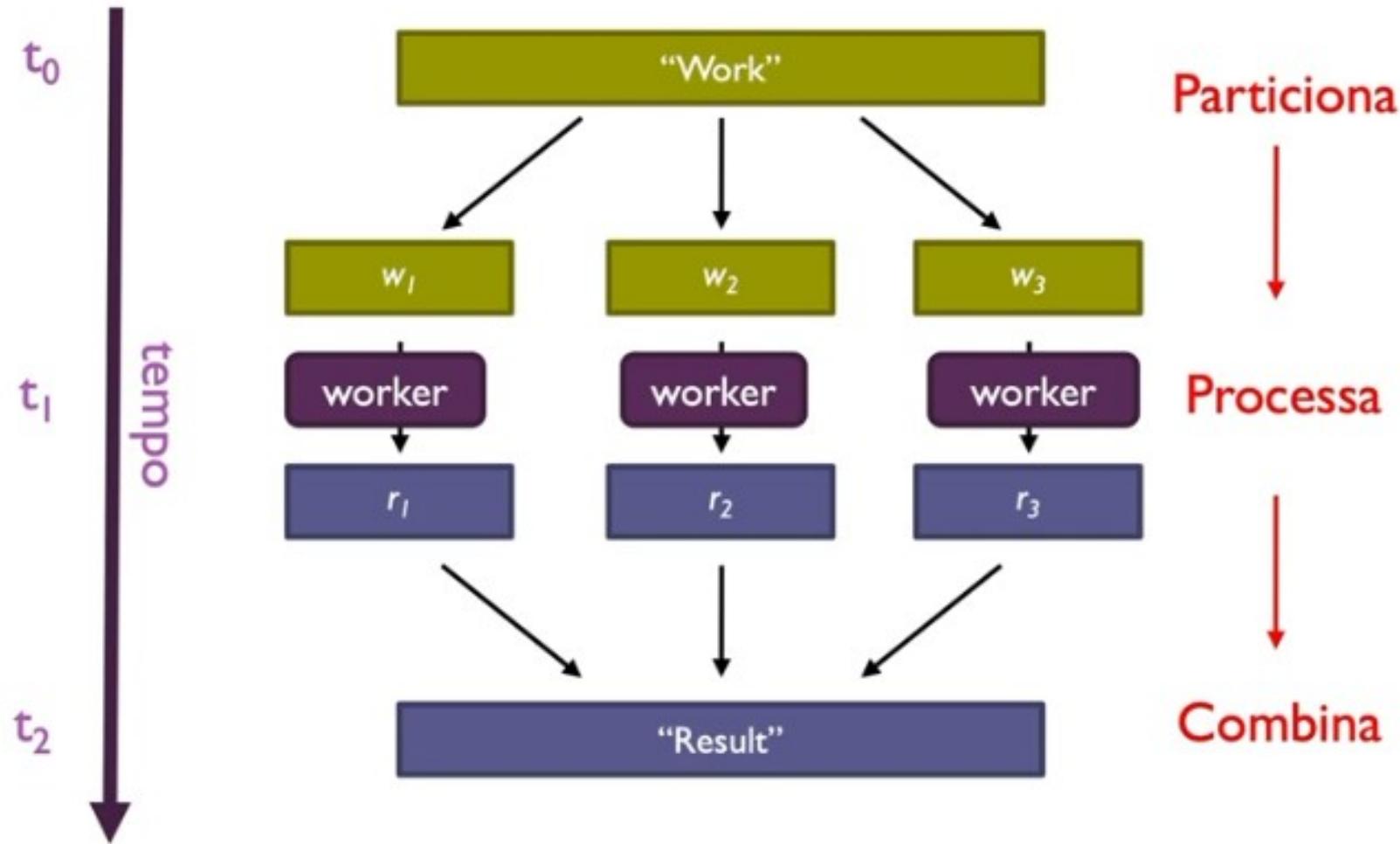
50

- Relembrando o algoritmo Merge Sort



Divisão e Conquista em Paralelo

51



Desafios da Paralelização

52

- Como distribuir unidades de trabalho entre workers?
- E se houverem mais unidades de trabalho do que workers?
- E se os workers precisarem compartilhar resultados parciais?
- Como agregar resultados parciais?
- Como saber se todos os workers terminaram?
- E se os workers “morrerem”?

Questão central: Sincronismo

53

- Problemas de paralelismo
 - Comunicação entre os workers
 - Acesso a recursos compartilhados
 - Em nosso caso, dados
- Assim, são necessários mecanismos de sincronização

Falta de Sincronismo

54



Gerenciando Múltiplos Workers

55

- Executam em uma ordem desconhecida
- Podem interromper um ao outro
- Podem precisar trocar resultados parciais
- Podem acessar dados compartilhados em qualquer ordem

Transações em BD

Esse conteúdo é visto em BD II, mas como o ensino remoto não permitiu que víssemos lá, vamos estudar agora!

Transação

57

- É um termo que se refere a uma **coleção de operações** que formam uma única unidade lógica de processamento de BD
- Inclui uma ou mais **operações de acesso** ao banco de dados
- Especifica-se os limites de uma transação com as instruções **begin transaction** e **end transaction**

Transação

58

- Tipos de transação:
 - **Somente de leitura**
 - **De leitura-gravação**
- As operações de acesso ao BD que uma transação pode fazer são:
 - **read_item(x)**: Lê um item do BD chamado X para uma variável do programa
 - **write_item(x)**: Grava o valor da variável de programa X no BD

Exemplos de transações

59

T_1

```
read_item(X);  
X := X-N  
write_item(X);  
read_item(Y);  
Y := Y+N;  
write_item(Y);
```

T_2

```
read_item(X);  
X := X+M  
write_item(X);
```

Transação

60

- As transações submetidas pelos diversos usuários podem ser executadas simultaneamente, **acessar e atualizar os mesmo itens do BD**
- Se essa execução simultânea for **descontrolada**, ela pode causar problemas:
 - O problema da atualização perdida
 - O problema da atualização temporária (ou leitura suja)
 - O problema do resumo incorreto
 - O problema da leitura não repetitiva

O problema da atualização perdida

61

T_1	T_2	
<pre>read_item(X); X := X-N</pre>	<pre>read_item(X); X := X+M</pre>	$X = R\$ 1000,00$ $N = R\$ 500,00$ $Y = R\$ 200,00$ $M = R\$ 300,00$
<pre>write_item(X); read_item(Y); Y := Y+N; write_item(Y);</pre>	<pre>write_item(X);</pre>	

O problema da atualização temporária (ou leitura suja)

62

T_1	T_2	
<pre>read_item(X); X := X-N write_item(X);</pre>		$X = R\$ 1000,00$ $N = R\$ 500,00$ $Y = R\$ 200,00$ $M = R\$ 300,00$
<pre>read_item(Y); Y := Y-M; write_item(Y);</pre>	<pre>read_item(X); X := X+M write_item(X);</pre>	

Tempo



O problema do resumo incorreto

63

T_1	T_2	
<pre>read_item(X); X := X-N write_item(X);</pre>	<pre>soma := 0 read_item(A) soma := soma + A ... read_item(X); soma := soma + X read_item(Y); soma := soma + Y</pre>	<p>Calcula o número total de reservas em todos os voos</p>
<pre>read_item(Y); Y := Y+N; write_item(Y);</pre>		<p>A = 150 X = 50 Y = 100 N = 20</p>

O problema da leitura não repetitiva

64

T_1	T_2
<code>read_item(X); verifica(X);</code>	<code>read_item(X); verifica(X);</code>
<code>read_item(X); reserva(X); write_item(X);</code>	<code>...</code> <code>read_item(X); reserva(X); write_item(X);</code>

A vertical arrow labeled "Tempo" points downwards, indicating the sequence of events over time.

Transação

65

- Na execução de uma transação o SGBD deve garantir:
 - Todas as operações na transação foram concluídas com sucesso e seu efeito será gravado permanentemente no BD (**confirmada - committed**)
 - Transação NÃO terá nenhum efeito sobre o BD ou outras transações (**abortada**)

O que faz uma transação falhar

66

- Computador falhar por hardware, software ou rede
- Erro durante execução de operação na transação:
divisão por zero
- Condições de exceção detectadas pela transação
(necessitam o cancelamento da mesma): saldo
insuficiente na conta
- Imposição de controle de concorrência
- Problemas físicos e catástrofes

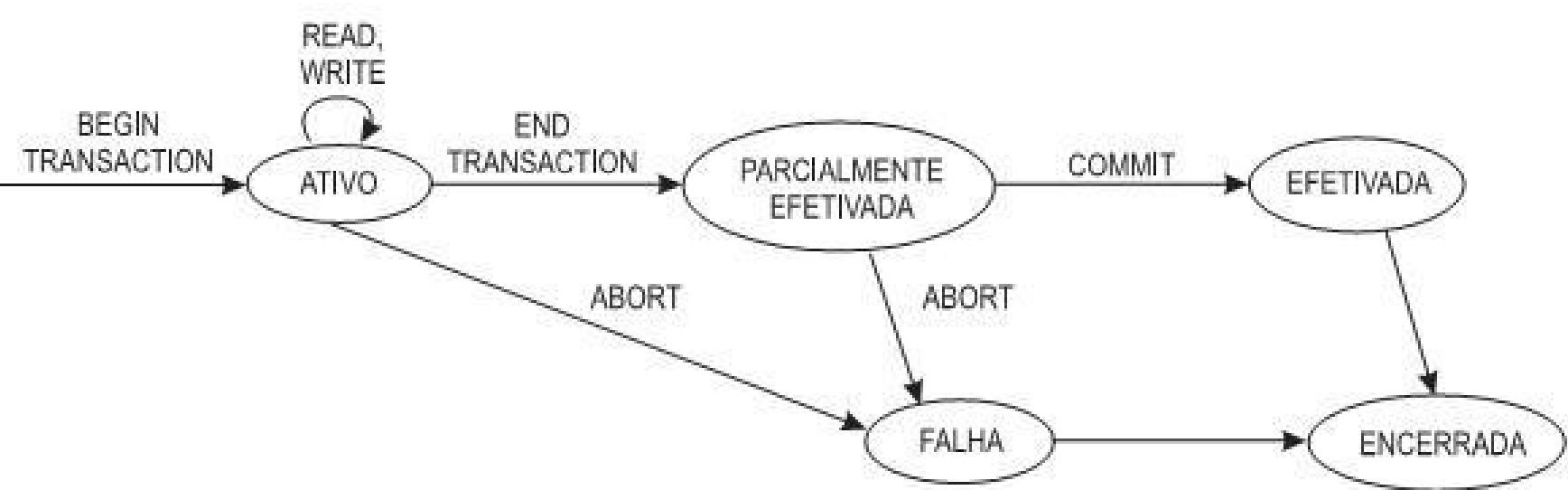
Estados de transação

67

- Existem as seguintes operações
 - BEGIN_TRANSACTION
 - READ ou WRITE
 - END_TRANSACTION
 - COMMIT_TRANSACTION
 - ROLLBACK (ou ABORT)

Diagrama de estados de uma transação

68



Propriedades desejáveis das transações

69

- **A**tomicidade
 - Deve ser realizada em sua totalidade ou não ser realizada de forma alguma
- **C**onsistência (Preservação)
 - Deve levar o BD de um estado consistente para outro
- **I**solamento
 - A execução de uma transação não deve ser interferida por quaisquer outras transações que acontecem simultaneamente
- **D**urabilidade (Persistência)
 - Após o ponto de confirmação, as alterações devem persistir no BD

Transação no MySQL

70

- Apenas o InnoDB suporta transações a nível ACID
- Operações
 - START TRANSACTION
 - ROLLBACK
 - COMMIT
- Ex:

```
START TRANSACTION;  
UPDATE jogadores SET idade = 26 WHERE codigo = 16;  
COMMIT;
```

Transação no MySQL + PHP

71

```
<?php  
$mysqli = new mysqli("127.0.0.1", "my_user", "my_password", "sakila");  
  
if ($mysqli->connect_errno) {  
    printf("Connect failed: %s\n", $mysqli->connect_error);  
    exit();  
}  
  
$mysqli->begin_transaction(MYSQLI_TRANS_START_READ_ONLY);  
  
$mysqli->query("SELECT first_name, last_name FROM actor");  
$mysqli->commit();  
  
$mysqli->close();  
?>
```

Fonte: www.php.net/manual/pt_BR/mysqli.begin-transaction.php

Exercícios

72

1. Considere:

- I. Se uma transação é concluída com sucesso (operação commit bem sucedida), então seus efeitos são persistentes.
- II. Ou todas as ações da transação acontecem, ou nenhuma delas acontece.

As propriedades (I) e (II) das transações em SGBDs, significam, respectivamente,

- a) durabilidade e consistência.
- b) persistência e automação.
- c) isolamento e atomicidade.
- d) durabilidade e atomicidade.
- e) consistência e persistência.

Exercícios

73

1. Considere:

- I. Se uma transação é concluída com sucesso (operação commit bem sucedida), então seus efeitos são persistentes.
- II. Ou todas as ações da transação acontecem, ou nenhuma delas acontece.

As propriedades (I) e (II) das transações em SGBDs, significam, respectivamente,

- a) durabilidade e consistência.
- b) persistência e automação.
- c) isolamento e atomicidade.
- d) durabilidade e atomicidade.
- e) consistência e persistência.

Exercícios

74

2. Que problema ocorre quando duas transações que acessam os mesmos itens de dados do banco de dados têm suas operações intercaladas, tornando com isso o valor de alguns itens do banco de dados incorretos?
- a) Atualização temporária.
 - b) Leitura suja.
 - c) Resumo incorreto.
 - d) Leitura não repetitiva.
 - e) Atualização perdida.

Exercícios

75

2. Que problema ocorre quando duas transações que acessam os mesmos itens de dados do banco de dados têm suas operações intercaladas, tornando com isso o valor de alguns itens do banco de dados incorretos?
- a) Atualização temporária.
 - b) Leitura suja.
 - c) Resumo incorreto.
 - d) Leitura não repetitiva.
 - e) Atualização perdida.

Exercícios

76

3. Em um banco de dados, uma transação é um conjunto de operações, delimitadas por um início e um fim. Independentemente da forma como a transação foi iniciada, esta sempre será finalizada por meio de dois comandos:
- I. o primeiro grava definitivamente os efeitos dos comandos de uma transação;
 - I. o segundo desfaz os efeitos dos comandos da transação.

Esses comandos são respectivamente:

- a)ABEND e SAVEPOINT.
- b)SAVEPOINT e COMMIT.
- c)COMMIT e ROLLBACK.
- d)ROLLBACK e CKECKPOINT.
- e)CHECKPOINT e ABEND.

Exercícios

77

3. Em um banco de dados, uma transação é um conjunto de operações, delimitadas por um início e um fim. Independentemente da forma como a transação foi iniciada, esta sempre será finalizada por meio de dois comandos:
- I. o primeiro grava definitivamente os efeitos dos comandos de uma transação;
 - I. o segundo desfaz os efeitos dos comandos da transação.

Esses comandos são respectivamente:

- a)ABEND e SAVEPOINT.
- b)SAVEPOINT e COMMIT.
- c)COMMIT e ROLLBACK.
- d)ROLLBACK e CKECKPOINT.
- e)CHECKPOINT e ABEND.

Exercícios

78

4. Uma transação de banco de dados deve possuir as propriedades abaixo, exceto uma. Assinale a alternativa que a apresenta.

- a) Durabilidade.
- b) Atomicidade.
- c) Confidencialidade.
- d) Consistência.
- e) Isolamento.

Exercícios

79

4. Uma transação de banco de dados deve possuir as propriedades abaixo, exceto uma. Assinale a alternativa que a apresenta.

- a) Durabilidade.
- b) Atomicidade.
- c) Confidencialidade.
- d) Consistência.
- e) Isolamento.

Problema Típico de Big Data

80

- Na prática, uma grande parte dos problemas não possuem grande volume de dependência entre os dados processados
 - Iterar sobre um grande número de registros
 - Extrair “algo” de interesse de cada registro
 - Distribuir e ordenar resultados intermediários
 - Agregar resultados intermediários
 - Gerar a saída final

Problema Típico de Big Data - Exemplos

81

- Seja uma coleção de milhões de documentos formados por palavras. Considere que cada documento é um registro
- Tarefas
 - Agregação: Quantas vezes cada palavra ocorre em uma coleção de documentos?
 - Indexação: Gerar uma lista de documentos em que cada palavra ocorre
 - Mineração de padrões: Gerar uma lista de palavras que mais co-ocorrem com cada palavra
 - Aprendizagem de máquina: Gerar um modelo de linguagem de N-Grams baseado na coleção de documentos

Modelos de Computação Paralela (CP) para Big Data

82

- Baseiam-se na existência de um **cluster computacional** composto por vários hosts similares e interligados por uma rede de comunicação de alto desempenho e tolerante a falhas
- Implementações destes modelos fornecem primitivas de programação que encapsulam os detalhes de processamento no cluster
- The datacenter as a computer!! (BARROSO, CLIDARAS e HÖLZLE, 2013)
- **MapReduce/Hadoop e Spark**

Modelos de CP para Big Data

83

- MapReduce
 - Proposto pela Google, publicado em 2004
 - Implementação proprietária em C++
 - Bindings em Java e Python
 - Hadoop: implementação open-source em Java
 - Desenvolvida pela Yahoo, agora um projeto Apache
 - Usado em produção: Yahoo, FB, Twitter, LinkedIn, Netflix...
 - Plataforma muito utilizada até 2012
 - Ecossistema de software amplo

Modelos de CP para Big Data

84

- Spark
 - Estende o MapReduce
 - Processamento de dados em memória principal
 - Execução de iterações
 - Trata naturalmente o processamento de fluxo de dados
 - Alto nível
 - Compatível com o ecossistema Hadoop
 - Compatível com vários sistemas de gerenciamento de clusters (Hadoop YARN, Apache Mesos, Kubernetes, etc)

Cluster Computacionais

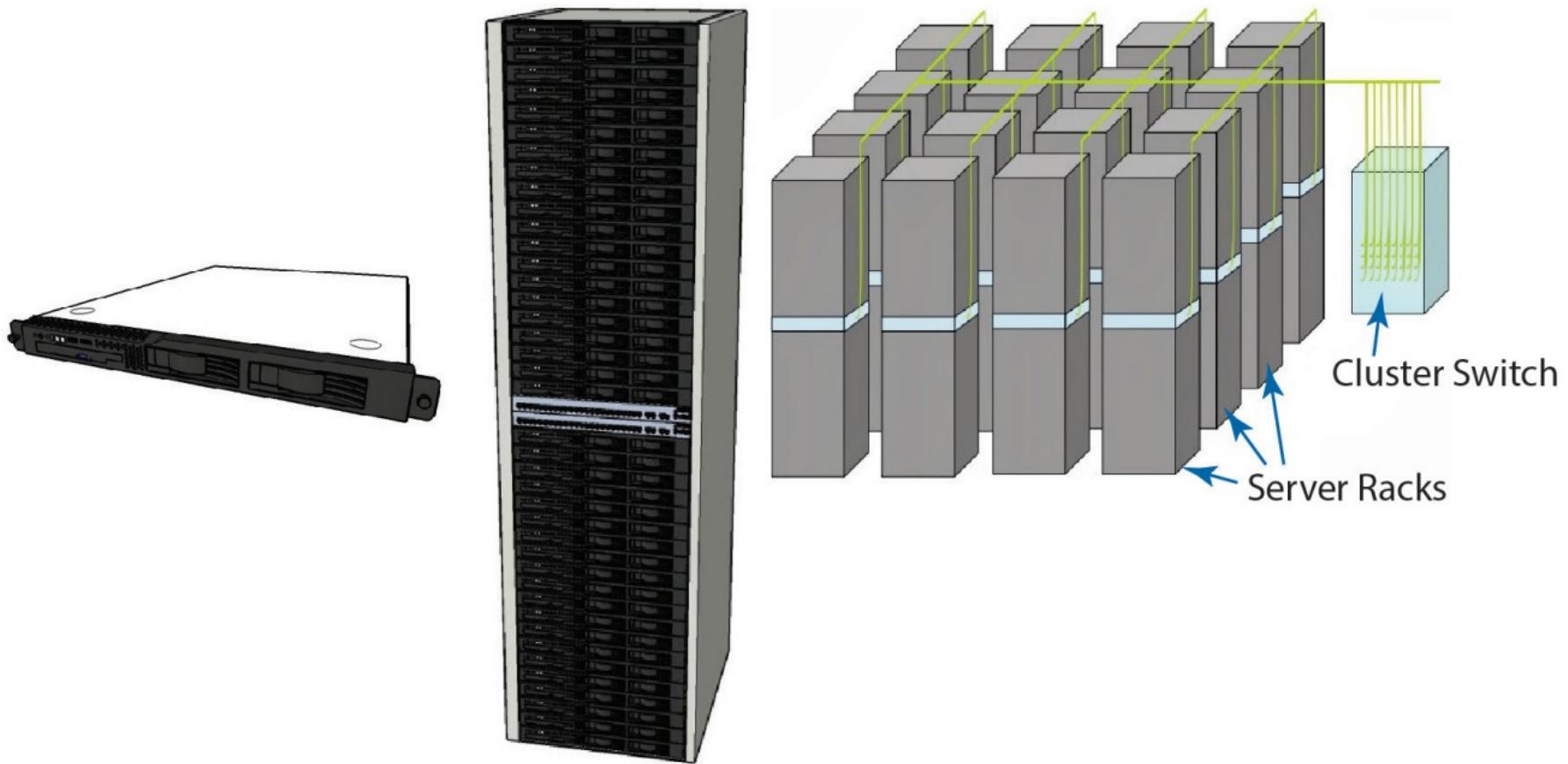
Cluster Computacionais

86

- Cluster pode ser traduzido como “agrupar” ou “agrupamento”
 - É o nome dado a um sistema que **relaciona vários computadores para que trabalhem de forma unificada** e, assim, entreguem um resultado ou objetivo
- Dificilmente uma empresa pequena ou média terá o próprio cluster
 - Utilizam serviços na nuvem

Componentes Básicos

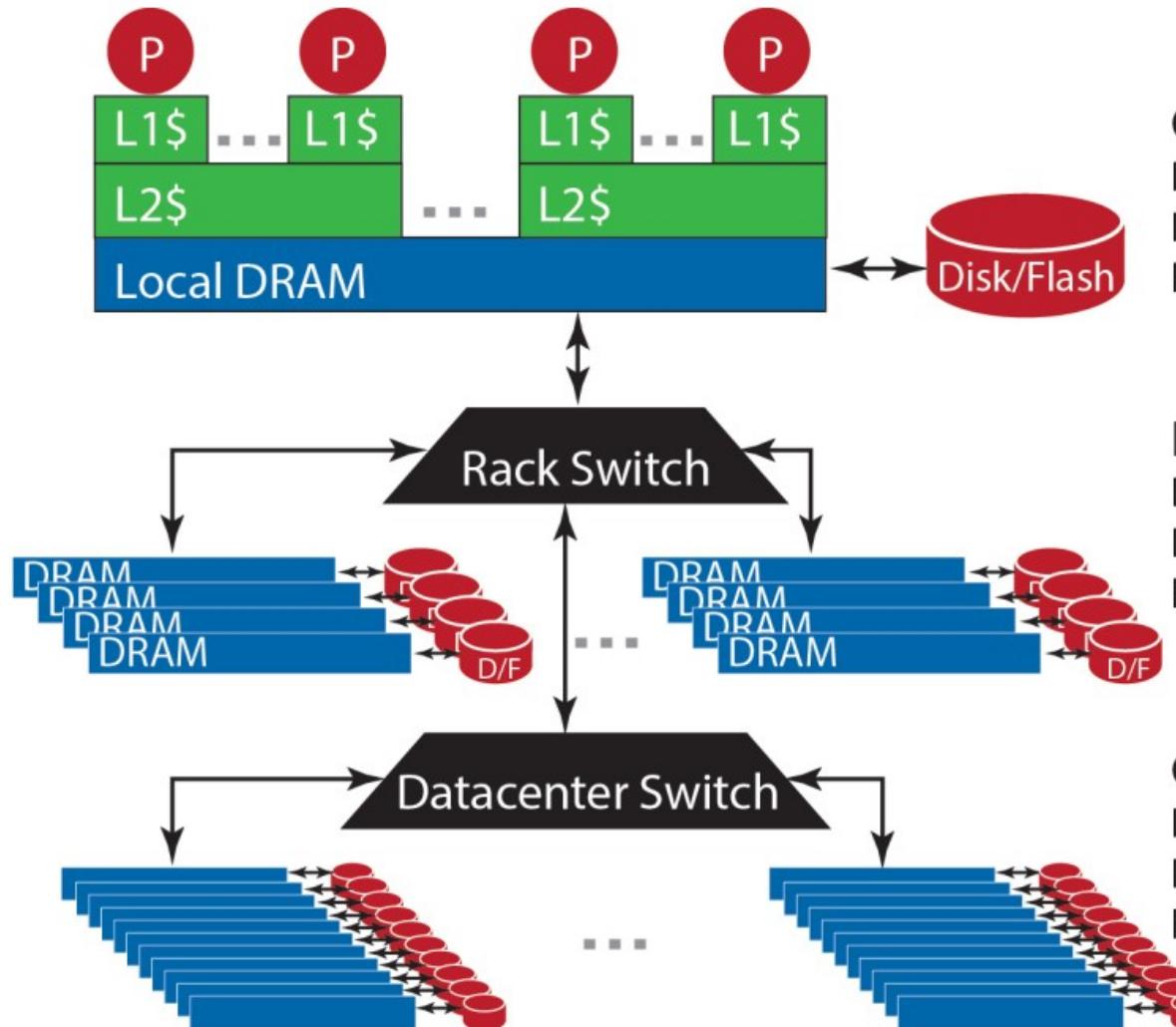
87



(BARROSO, CLIDARAS e HÖLZLE, 2013)

Hierarquia de Armazenamento

88



One Server

DRAM: 16 GB, 100 ns, 20 GB/s
Disk: 2T B, 10 ms, 200 MB/s
Flash: 128 GB, 100 us, 1 GB/s

Local Rack (80 servers)

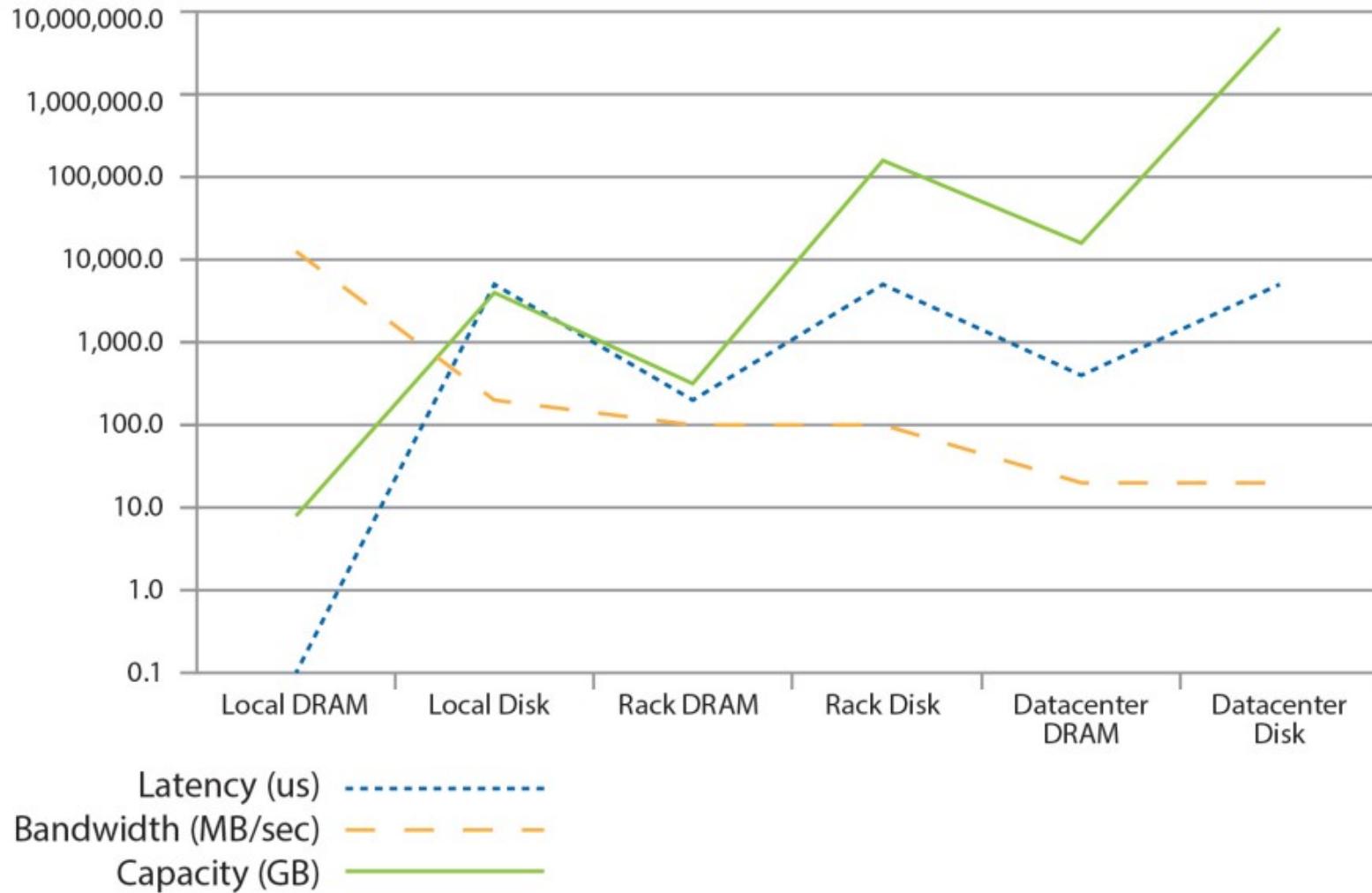
DRAM: 1 TB, 300 us, 100 MB/s
Disk: 160 TB, 11 ms, 100 MB/s
Flash: 20 TB, 400 us, 100 MB/s

Cluster (30 racks)

DRAM: 30 TB, 500 us, 10 MB/s
Disk: 4.80 PB, 12 ms, 10 MB/s
Flash: 600 TB, 600 us, 10 MB/s

Hierarquia de Armazenamento

89



(BARROSO, CLIDARAS e HÖLZLE, 2013)

“Big” Ideias

90

- Scale “out”, not “up”
 - Dimensionar ao invés de expandir
- Mover o “processamento” para os “dados”, e não o contrário
- Acessar dados sequencialmente
 - Evitar acesso aleatório
- Escalabilidade deve ser simples
- Assumir que há falhas

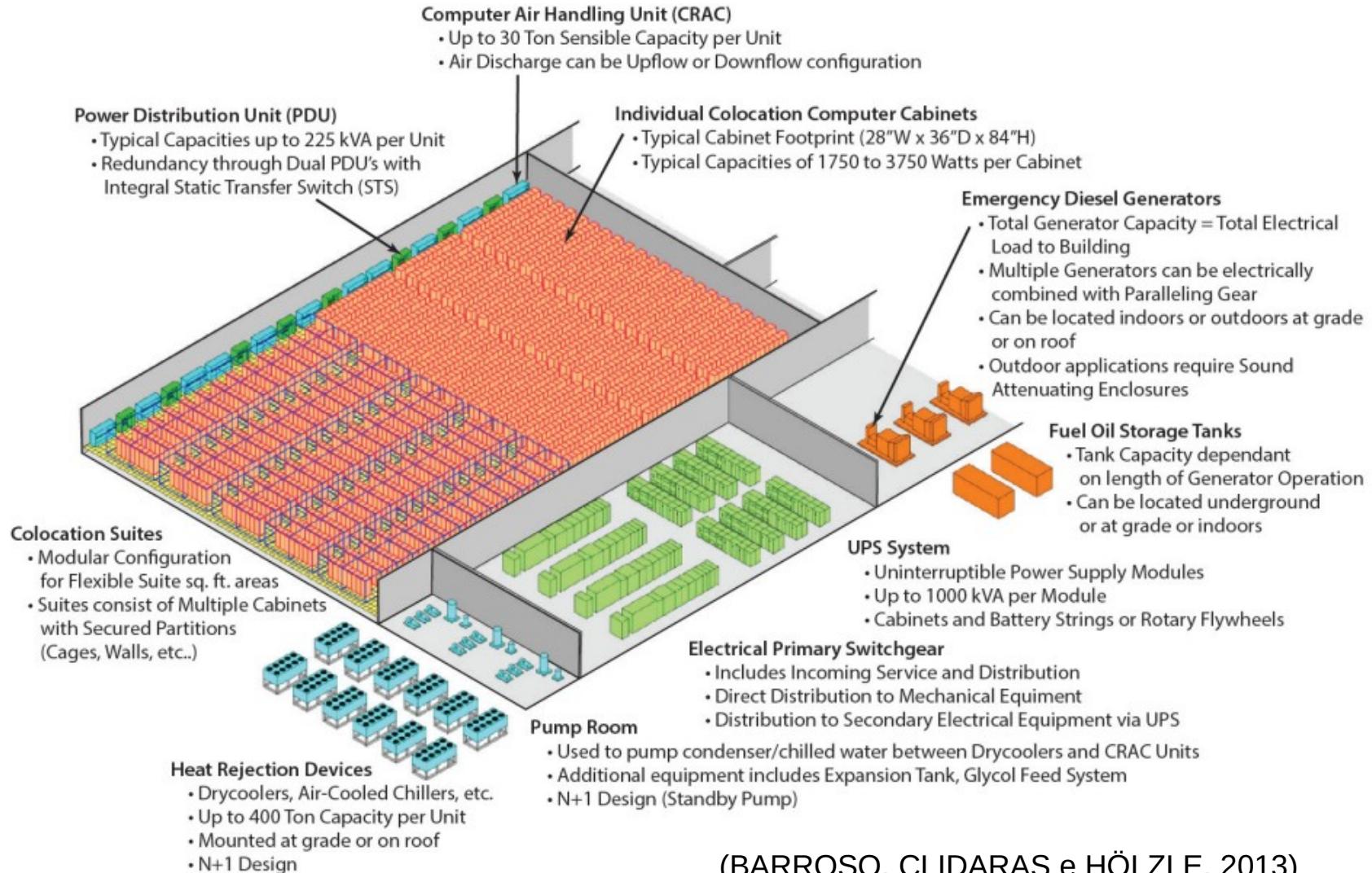
Acesso sequencial vs. randômico

91

- BD de 1TB com registros de 100 bytes
 - Vamos atualizar 1% dos registros
- Cenário 1: acesso randômico
 - Cada atualização leva ~30ms (seek, read, write)
 - 10^8 atualizações = ~35 dias
- Cenário 2: reescrever todos os registros
 - Assuma 100 MB/s de taxa de transferência
 - Tempo = 5,6 horas

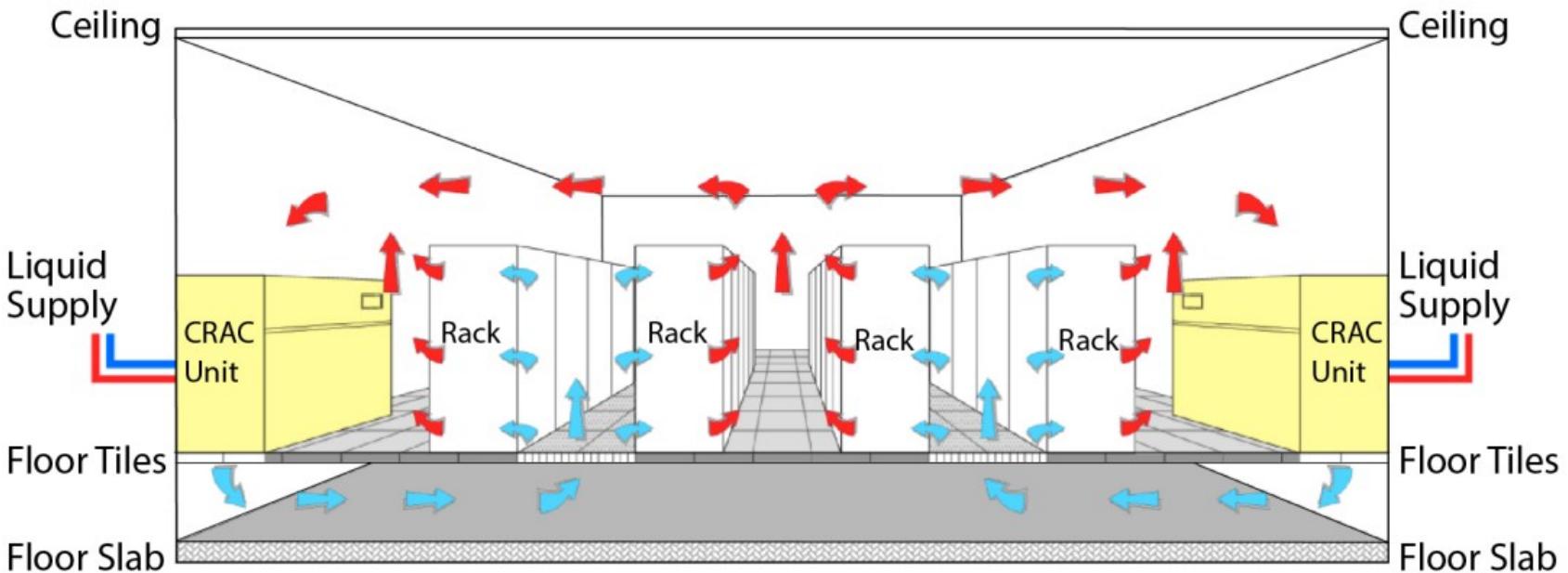
Anatomia de um Datacenter

92



Anatomia de um Datacenter

93



(BARROSO, CLIDARAS e HÖLZLE, 2013)

Datacenters Reais

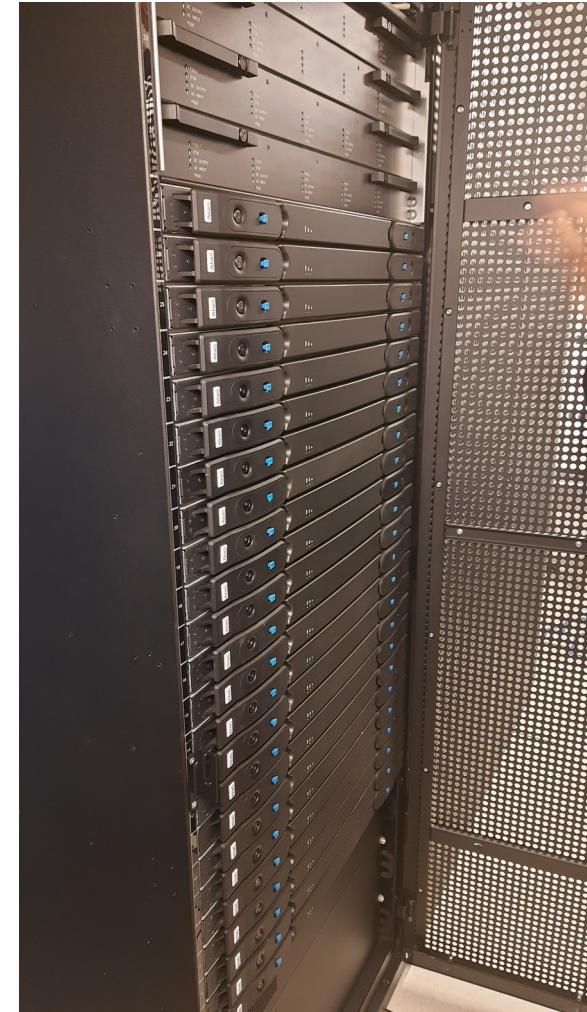
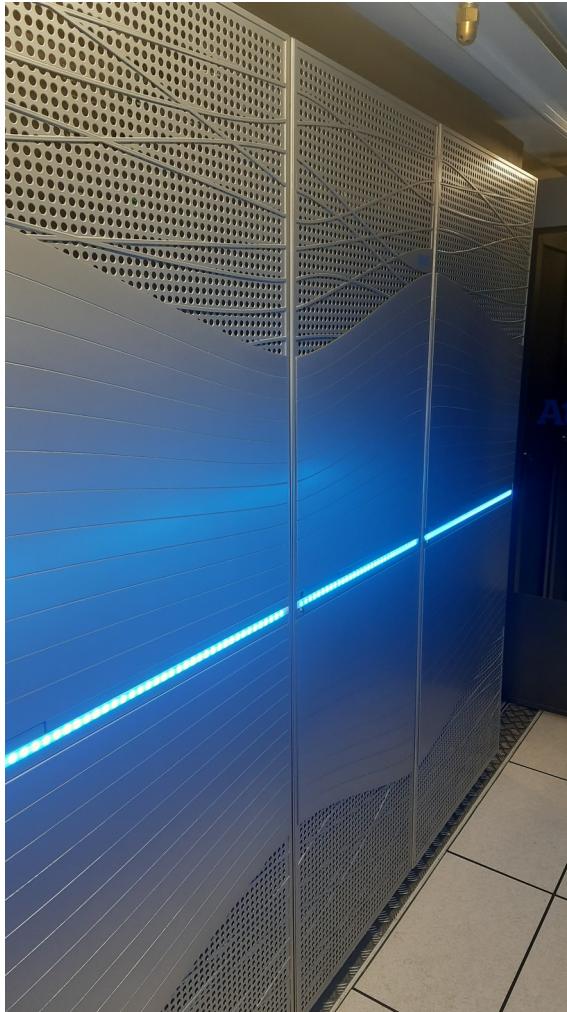
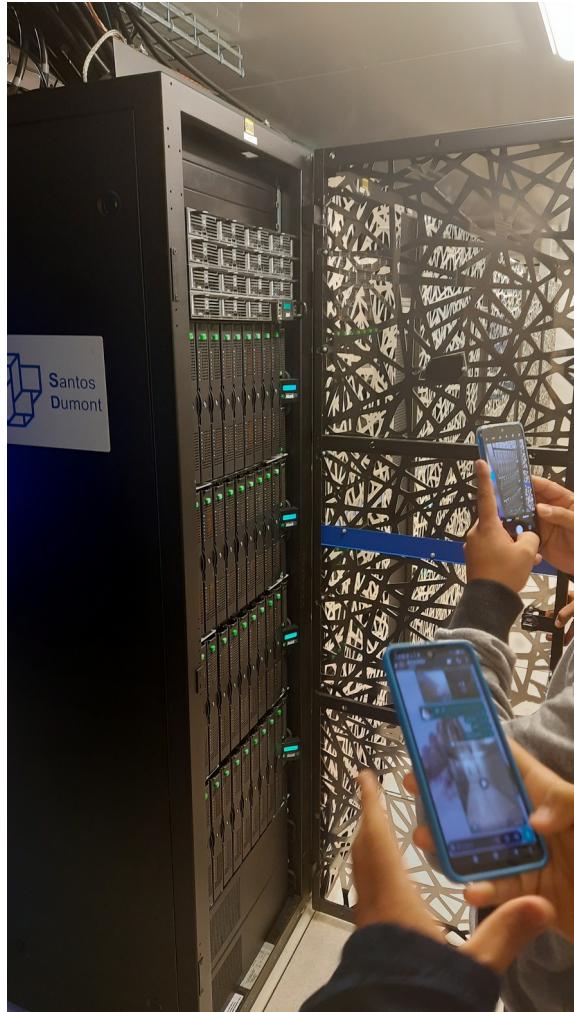
Santos Dumont

95



Santos Dumont

96



Santos Dumont

97

- Localizado na sede do Laboratório Nacional de Computação Científica (LNCC), em Petrópolis-RJ
- Possui um total de 36.472 núcleos de CPU, distribuídos em 1.134 nós computacionais
- É dotado de um nó diferenciado com número elevado de núcleos (240) e arquitetura de memória compartilhada de grande capacidade (6 Tb em um único espaço de endereçamento)
- Existe um nó especialmente projetado para aplicações de Inteligencia Artificial (Deep Learning) que dispõe de 8 GPUs NVIDIA Tesla V100-16Gb com Nvlink, totalizando 40.960 CUDA-core e 5120 Tensor-core
- Possui 1.7PB de capacidade

Amazon

98



Google

99



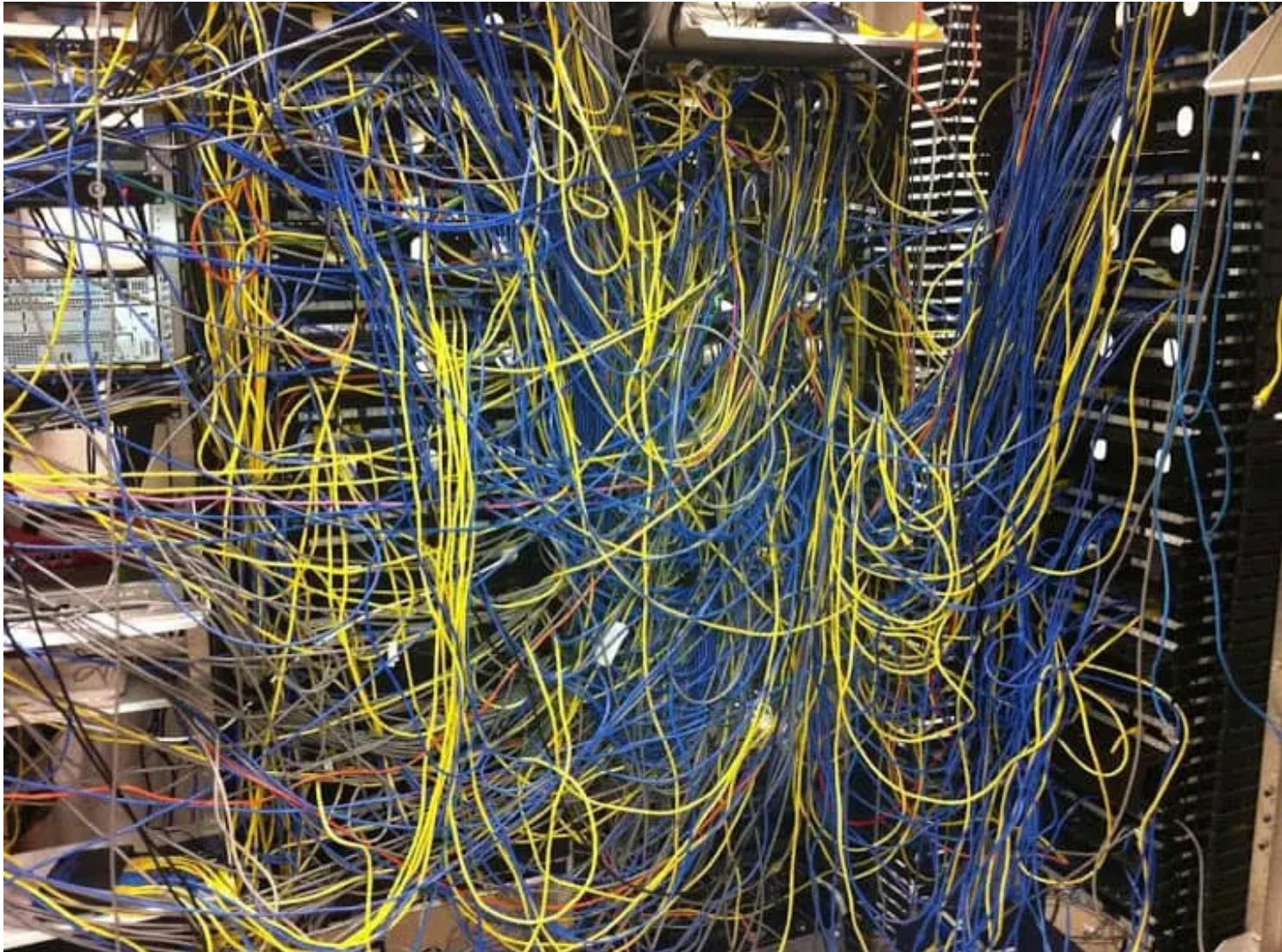
Facebook

100



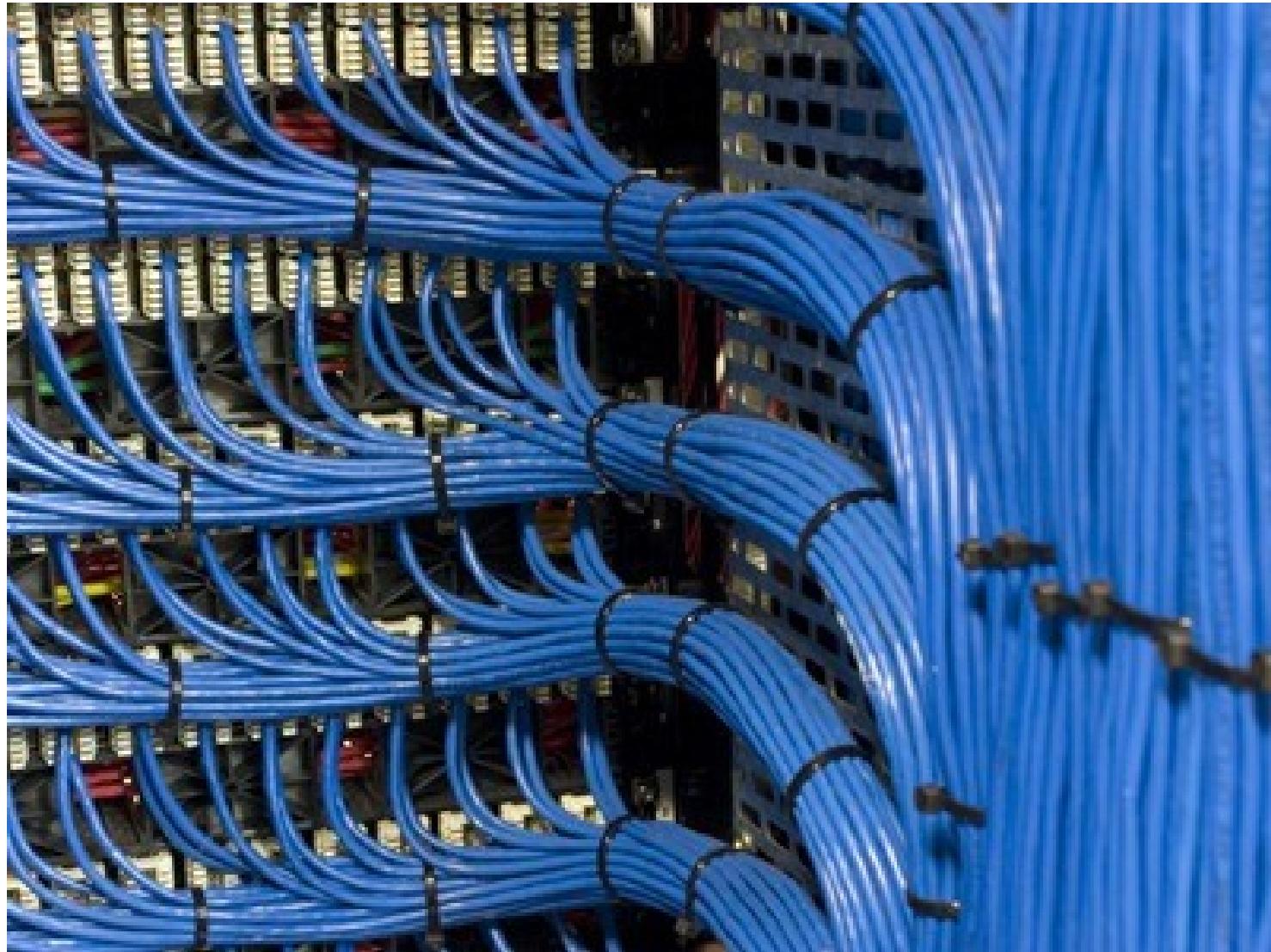
Imagina os cabos

101



Imagina os cabos

102



Computação nos mares

103



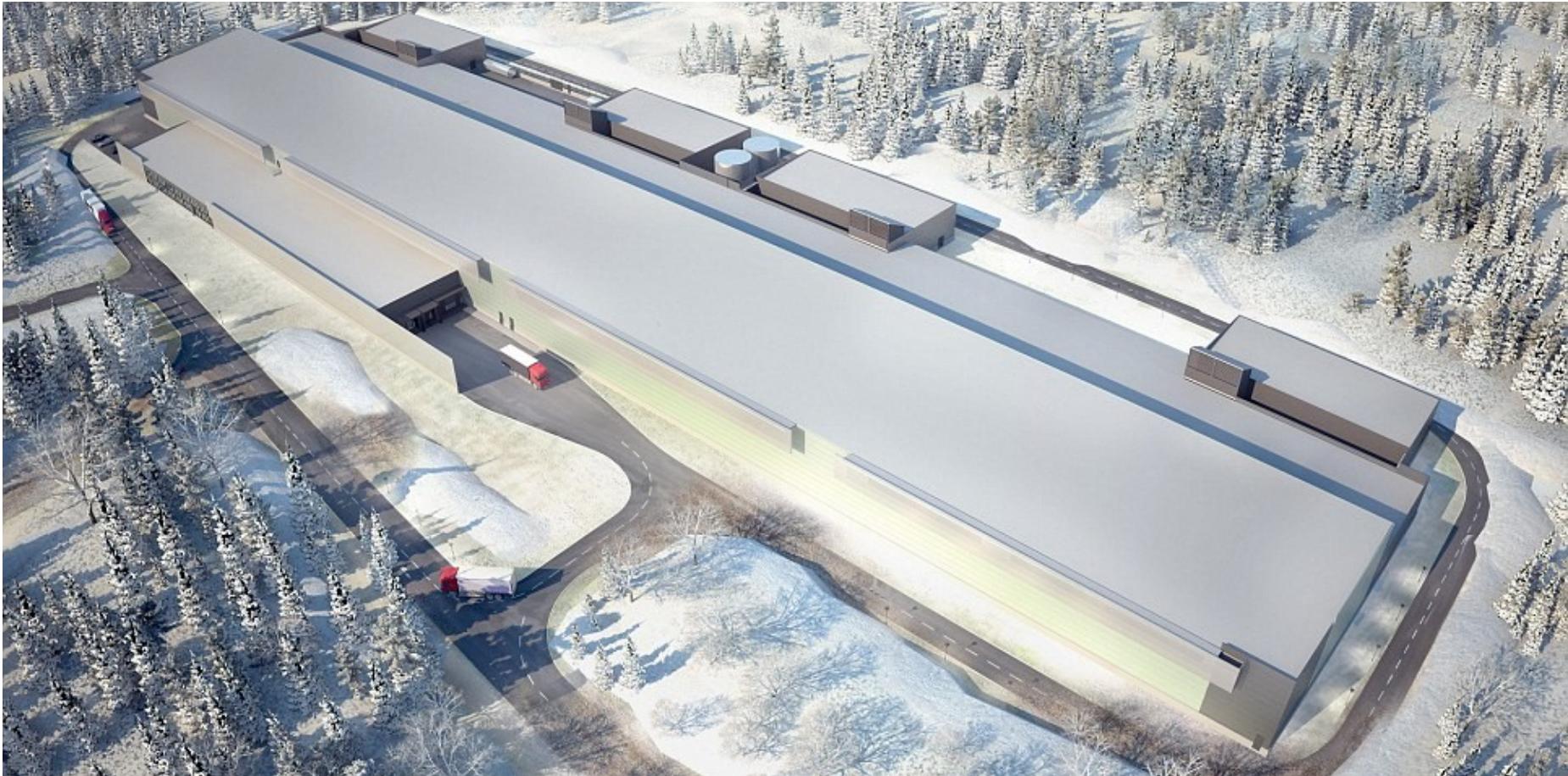
Computação nos mares

104



Datacenter no Ártico

105



Referências

106

- › BANKO, Michele; BRILL, Eric. Scaling to very very large corpora for natural language disambiguation. In: Proceedings of the 39th annual meeting of the Association for Computational Linguistics. 2001. p. 26-33.
- › BARROSO, Luiz André; CLIDARAS, Jimmy; HÖLZLE, Urs. The datacenter as a computer: An introduction to the design of warehouse-scale machines. *Synthesis lectures on computer architecture*, v. 8, n. 3, p. 1-154, 2013.
- › HALEVY, Alon; NORVIG, Peter; PEREIRA, Fernando. The unreasonable effectiveness of data. *IEEE intelligent systems*, v. 24, n. 2, p. 8-12, 2009.

Dúvidas?

107



jean.camara@ifsudestemg.edu.br