

TSW017

Grandes Sistemas de Dados Projeto e Gerenciamento

Prof. Dr. Jorge Rady

1

Big Data Motivação e Definições

2

Big Data - Motivação

- Aumento da geração e armazenamento digital de dados
- Internet das Coisas (Internet of Things), personalização de serviços, novas fontes de dados
- Mídias sociais, meteorologia, clickstream, e-mails, dados geo-espaciais, imagens médicas, . . .
- Aumento da capacidade de armazenamento
- Aumento da capacidade de processamento em geral: Computadores, Tablets, Smartphones, . . .

3

Big Data - Motivação

- Dados contêm informação/conhecimento significativo e que pelo seu custo e tempo de obtenção merecem tratamento adequado
- Compreensão: análise, captura, tratamento, armazenamento, compartilhamento, consulta e visualização
- Problema: tomar grandes e complexos conjuntos de dados que os aplicativos tradicionais não consigam processar em tempo adequado

4

Big Data - Primeira Citação

Application-Controlled Demand Paging for Out-of-Core Visualization

Michael Cox and David Ellsworth¹
Report NAS-97-010, July 1997

NASA Ames Research Center
MS T27A-2
Moffett Field, CA 94035-1000

1 Introduction

Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of *big data*. When data sets do not fit in main memory (*in core*), or when they do not fit even on local disk, the most common solution is to acquire more resources. This *write-a-check* algorithm

5

Big Data

- ❖ Geração de tecnologias e arquiteturas que visam extrair valor a partir de bases de dados muito grandes, com grande variedade de origens, com grande velocidade de captura e análise
- ❖ “Big” no Big Data não é somente o tamanho da base de dados
- ❖ Não se refere a nenhum patamar de volume de dados
- ❖ Os dados são “big” quando seu tamanho, variedade e velocidade de geração tornam-se elementos ativos do problema

6

Big Data

The DAMA Guide to
The Data Management
Body of Knowledge
(DAMA-DMBOK Guide)

1.1 Data: An Enterprise Asset

Data and information are the lifeblood of the 21st century economy. In the Information Age, data is recognized as a vital enterprise asset.

“Organizations that do not understand the overwhelming importance of managing data and information as tangible assets in the new economy will not survive.”

Tom Peters, 2001

7

Big Data



Big Data consists of extensive datasets—primarily in the characteristics of volume, variety, velocity, and/or variability—that require a scalable architecture for efficient storage, manipulation, and analysis.

The **Big Data paradigm** consists of the distribution of data systems across horizontally coupled, independent resources to achieve the scalability needed for the efficient processing of extensive datasets.

Big Data engineering includes advanced techniques that harness independent resources for building scalable data systems when the characteristics of the datasets require new architectures for efficient storage, manipulation, and analysis.

8

Big Data - Definição



Termo usado para descrever dados com grande volume, velocidade e/ou variedade que requer novas tecnologias e técnicas para capturar, armazenar e analisar tais dados. Usado para aprimorar o processo de tomada de decisões e otimizar processos.

9

Big Data - Definição



Big Data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big Data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a single data set.

10

Big Data

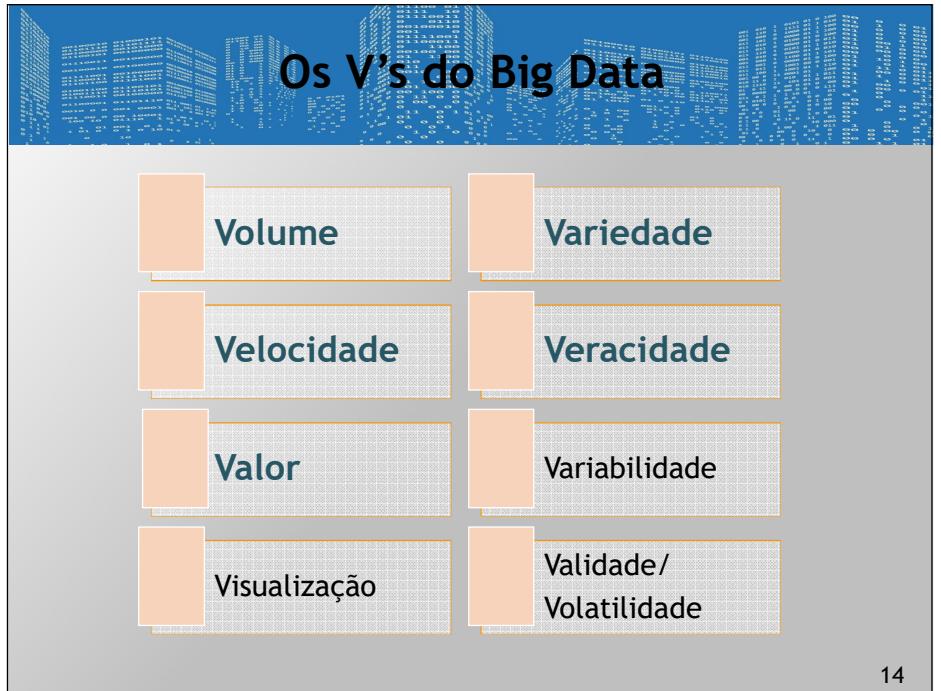
- ❖ Coleção de dados tão grande e complexa que torna difícil o processamento por meio de SGBD tradicionais
- ❖ Próprio quando for mais barato (simples) manter todos os dados do que ter um esforço para decidir o que eliminar (David Brower)

11

Big Data - Desafios

- ❖ Dados
 - Volume, variedade, velocidade, variabilidade, veracidade, visualização, valor
- ❖ Processo
 - Aquisição e armazenamento de dados
 - Limpeza de dados
 - Data Mining e Análise
- ❖ Gerenciamento
 - Privacidade, segurança, governança de dados, custos, propriedade dos dados

12



Big Data - 5 V's

Volume

- Quantidade crescente de pessoas e dispositivos conectados à Internet
- Volume de dados em constante aumento: terabytes → petabytes → exabytes → zetabytes . . .
- Geração constante
- Exponencialmente crescente
- Estatísticas?

15

Big Data - 5 V's

Variedade

- Múltiplos repositórios de origem
- Múltiplos formatos
- Estruturados, Não Estruturados, Semi Estruturados
- Textos, vídeos, redes sociais, tabelas, imagens, sensores, . . .

16

Dados Estruturados x Não Estruturados

Dados estruturados - SGBD Relacionais

- Possuem esquema: saber a priori o que se pode armazenar

Dados não estruturados - estruturas livres (sem esquema)

- Armazenar o que for necessário
- Pode-se adaptar o armazenamento: adicionar novos elementos ou deixar de armazenar um tipo de dados não mais necessário
- Cada item pode ter um conjunto distinto de tipos de dados

17

Big Data - 5 V's



Velocidade

- Geração de dados cada vez mais acelerada
- Necessidade de processamento mais rápido
- Data streaming: transferência de dados a altas taxas de transmissão/ recepção
- Dados em Batch → Periódicos → Near real-time → Real-Time
- Redes de sensores: veículos, cidades inteligentes, . . .
- Carros modernos: mais de 100 sensores que monitoram combustível, motor, pneus, ...
- Haverá 18.9 bilhões de conexões de rede - até 2020 (quase 2.5 conexões por pessoa na Terra)

18

Big Data - 5 V's



Veracidade

- Origem conhecida e confiável
- Confiança no significado e no conteúdo dos dados
- Qualidade dos dados: completeza, acurácia, . . .
- Qualidade dos dados: custo



Valor

- Grau de utilidade
- Retorno de investimento
- Mostrar grau aceitável / considerável de benefícios

19

Big Data - Variabilidade e Visualização



Variabilidade:

Alteração ou inserção de novos significados dos dados



Visualização:

utilizar a formas gráficas mais úteis e acessíveis



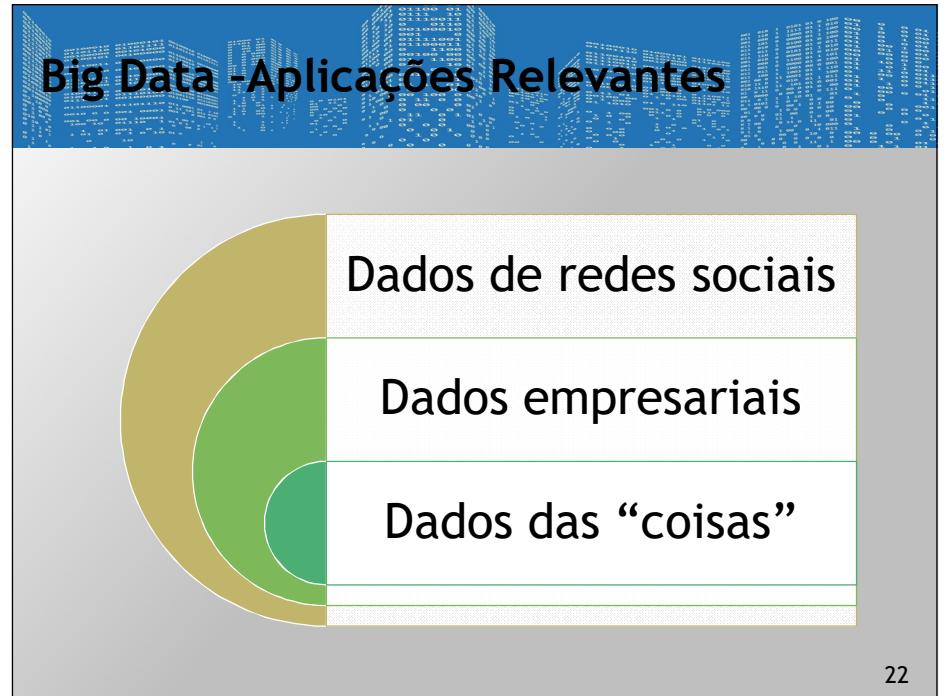
Validade / Volatilidade:

qual o período de utilidade dos dados

20



21



22

-
- A slide titled "Big Data - Aplicações Relevantes" set against a background of binary code forming a city skyline.
- **Geral:** combinação de múltiplas fontes de dados (E, SE, NE)
 - **Ciências:** física, pesquisas biomédicas,
 - **Instituições financeiras:** entender e aumentar satisfação de clientes; minimizar riscos e fraudes
 - **Governo:** gestão de serviços públicos, questões de defesa nacional,
 - **Saúde:** análise de registros de pacientes para predição/reação a eventos clínicos → tratamentos melhores e mais rápidos

23

-
- A slide titled "Big Data - Aplicações Relevantes" set against a background of binary code forming a city skyline.
- ❖ **Comércio:** Análise de Fidelidade, Análise da Concorrência e de Risco de Clientes (Pão de Açucar, Vivo, Amazon,)
 - ❖ **Análise do clima:** dados de centrais meteorológicas
 - ❖ **Análise de novos funcionários:** coleta e observação de dados em redes sociais
 - ❖ **Manutenção de equipamentos:** previsão de panes
 - ❖ **Educação:** identificar alunos com dificuldades, implementar melhores sistemas de avaliação

24

Big Data - Aplicações Relevantes

- ❖ Kroger - rede de varejo americana e Sprint - telecomunicações
 - Análise de fidelidade de clientes e rentabilidade
- ❖ UPS e Tesla
 - Cruzar dados de sensores de veículos: otimização da frota, de rotas, economia de combustível
 - Aprimorar desempenho e manutenção de veículos

25

Big Data - Aplicações Relevantes

COMPUTERWORLD

- ❖ Ford
 - Analisar padrões de direção
 - Redução de custos dos motoristas com seguros
- ❖ LHC (Large Hadron Collider): geração de dados chegou a 25 PB (2012)
- ❖ Observatório Apache Point (Novo México) gera 200 GB por noite, já tendo 140 TB de dados
- ❖ Governo Americano: análise de redes sociais, telefônicas, de dados na busca por padrões
- ❖ Netflix, Spotify, Waze, Yahoo, . . .

26

Big Data - Histórico

27

Big Data - Histórico

Evolução no volume de dados

- MB → GB : bases de dados
- GB → TB : paralelização no processamento e armazenamento de dados
- TB → PB
 - Internet
 - Google File System (GFS) e o Map Reduce
 - Múltiplos sensores
 - Bases NoSQL
- PB → EB : Projetos de Big Data (EMC, Oracle, Microsoft, Google, Amazon, IBM, ...)

28

Big Data - Histórico

Fatos históricos

- Termo utilizado no censo americano de 1880: o volume de dados coletado levou 7 anos para ser processado e mostrar seus resultados (Glatt, 2014)
- Evolução do significado “big data” a partir da década de 1970/80 MB → GB → TB . . .
- 2008: revista Nature publica edição especial sobre Big Data
- 2010: 15 artigos em periódicos científicos citando o termo → 380 em 2013
- 2011: revista Science publica edição especial sobre o processamento de Big Data

29

Big Data - Histórico

Fatos históricos

2012: Forum de Davos, Suíça declara o Big Data como um novo tipo de valor econômico

2012: Japão declara o desenvolvimento de Big Data como estratégia nacional

2012: Nações Unidas emitem o relatório Big Data for Development

30

Big Data - Aproveitando Oportunidades

BIG DATA:
SEIZING OPPORTUNITIES,
PRESERVING VALUES

Executive Office of the President



MAY 2014

31

Fatores considerados

Big Data e IoT: reunir economia industrial e economia da informação → redução do custo de manutenção e aumento da segurança

Redução de custos da saúde pela detecção de fraudes

Auxílio na guerra do Afeganistão, identificando condições perigosas aos americanos

Identificação de recém-nascidos com tendência a contraírem infecções

32

Perigos a serem evitados

Ferramenta usada como forma de discriminação de cidadãos

Possibilidade de obter dados privados dos cidadãos aparentemente ocultos (data fusion)



33

Recomendações

Preservar dados privativos: proteger as informações pessoais

Utilização preferencial na educação

Prevenir novos modos de discriminação

Assegurar o uso dentro da lei e da segurança nacional

Dados são um recurso público

34

Big Data - BDSSG

THE FEDERAL BIG DATA
RESEARCH AND DEVELOPMENT
STRATEGIC PLAN

THE NETWORKING AND INFORMATION
TECHNOLOGY RESEARCH AND
DEVELOPMENT PROGRAM



MAY 2016

35

The Big Data
Senior Steering
Group (BDSSG)

White House
Big Data R&D
Initiative

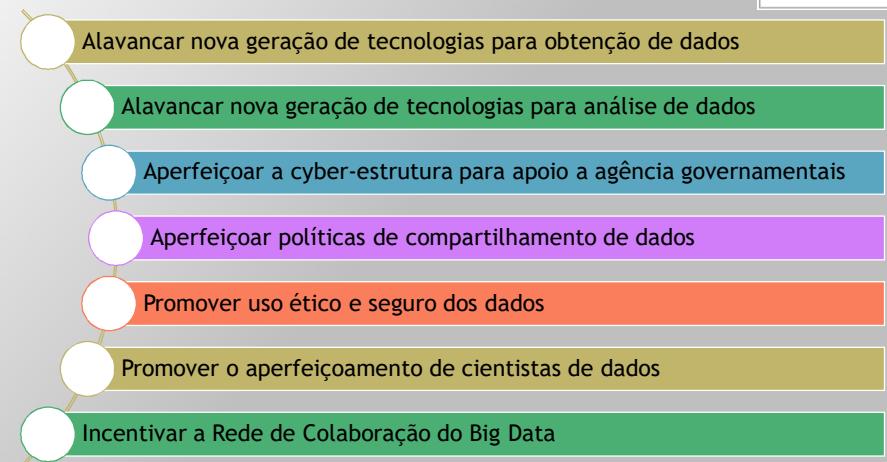
- Criação em 2011 sob o programa NITRD (Networking and Information Technology R&D)
- Identificar pesquisas/desenvolvimentos em Big Data

- Criação em 2012
- Abrangeu o BDSSG



36

Big Data - BDSSG - 7 Estratégias



37

Big Data - BDSSG - Áreas Prioritárias



Saúde

Ambiente e sustentabilidade

Resposta a emergências

Resiliência a desastres

Ciências

Educação

Cyber espaço seguro

Transporte e energia

Manufatura e robótica

38

Big Data - NIST 7 Volumes



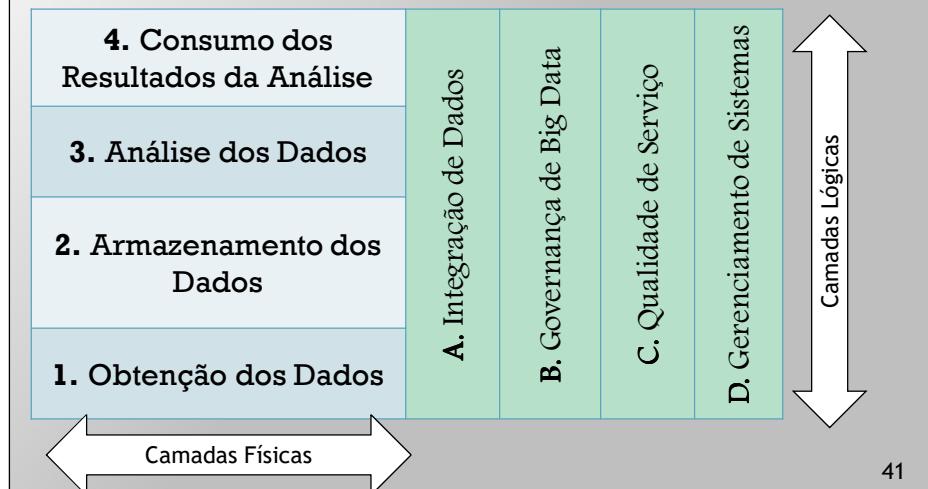
- Volume 1 Definitions
- Volume 2 Taxonomies
- Volume 3 Use Cases and General Requirements
- Volume 4: Security and Privacy
- Volume 5: Architectures White Paper Survey
- Volume 6: Reference Architecture
- Volume 7: Standards Roadmap

39

Camadas da Arquitetura Big Data

40

Camadas da Arquitetura Big Data



1. Obtenção dos Dados

- Obtenção de dados estruturados, semi-estruturados e não estruturados → Variedade
- Formatos variáveis
- Velocidade
- Frequência de obtenção dos dados: sob demanda, contínua, real-time
- Sensores, mídias sociais, imagens, . . .

42

2. Armazenamento dos Dados

Armazenar e/ou processar esses dados → cluster master / slave

3. Análise dos Dados

Utilizar ferramentas apropriadas (data mining, estatística, IA)

Gerar os documentos necessários

Entregar aos profissionais adequados

Analisar: métodos descritivo, diagnóstico, preditivo, prescritivo

43

4. Consumo dos Resultados da Análise

Humanos ou aplicativos

Utilização adequada desses resultados → tomada de decisões e de ações → visa melhorar a tomada de decisões nos negócios

44

A. Integração dos Dados

Definir as ferramentas de software para as camadas lógicas

Definir os tipos de bancos de dados a serem formados

45

B. Governança de Big Data

Gerenciar grandes volumes de dados

Definir as políticas de uso, armazenamento e remoção de dados

Reducir atritos operacionais

Proteger necessidades de stakeholders

Adotar mesmo entendimento dos dados

Adotar processos padrão

Reducir custos e aumentar a eficiência

Garantir transparência dos processos

46

C. Qualidade de Serviço

- Disponibilizar os dados em tempos adequados
- Verificar a correção e precisão dos dados
- Respeitar a política de privacidade

D. Gerenciamento de Sistemas

- Verificar logs
- Monitorar sempre
- Atender aos contratos com terceiros
- Manter backups

47

Big Data Tipos de análise

48

Big Data Analítico

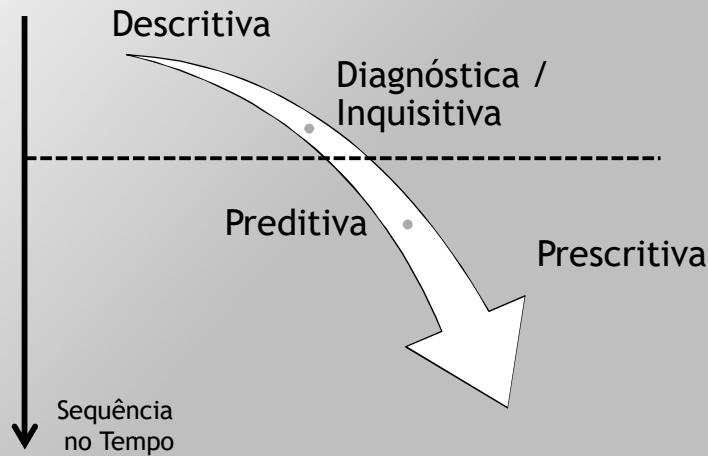
Processo de examinar grandes conjuntos de dados para obter padrões, correlações, tendências “ocultas”

Ferramentas: data mining, text analytics, análise estatística e preditiva



49

Big Data Analítico - Tipos de Análise



51

Big Data Analítico - Técnicas

Aprendizado de máquina: uso de algoritmos para reconhecer padrões complexos

Processamento de Linguagem Natural - IA e linguística: algoritmos para analisar a linguagem humana

Reconhecimento de Padrões: técnicas de aprendizado de máquina

50

Big Data Analítico - Tipos de Análise

Descritiva: o que aconteceu? Qual a sua frequência?

- Identificar e avaliar os atributos
- Estimar a contribuição dos atributos no resultado

Diagnóstica / Inquisitiva: porque aconteceu?

- Extrair padrões a partir dos dados

Preditiva: o que pode acontecer

- Determinar a probabilidade de possíveis resultados

Prescritiva: o que se recomenda fazer

- Ações indicadas para obter os resultados desejados

52

Big Data Analítico - Tipos de Análise

Conhecimentos importantes

Identificação de classes, associações, correlações, . . .

Tipping Point

Valor de um atributo que provoca alterações no comportamento do sistema sob análise

53

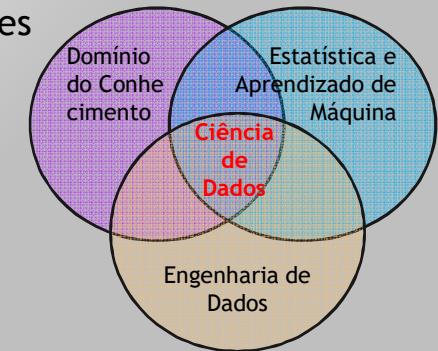
Big Data - Ética

55

Ciência de Dados e a Análise



- ❖ **Ciência de Dados:** extração de conhecimento diretamente dos dados por meio de um processo de descoberta ou de formulação e teste de hipóteses



54

Big Data - Ética

Exemplos

- Seguradoras determinar o valor do prêmio com base no histórico de locais frequentados pelo motorista
- Operadoras de planos de saúde fazer o mesmo de acordo com o histórico de compras de medicamentos
- Uso de informações de modo de vida (histórico de compra, lojas frequentadas, trajetos, dados financeiros) para induzir/influenciar futuras compras
- Rede de lojas americana Target, visando determinar as mulheres provavelmente grávidas

56

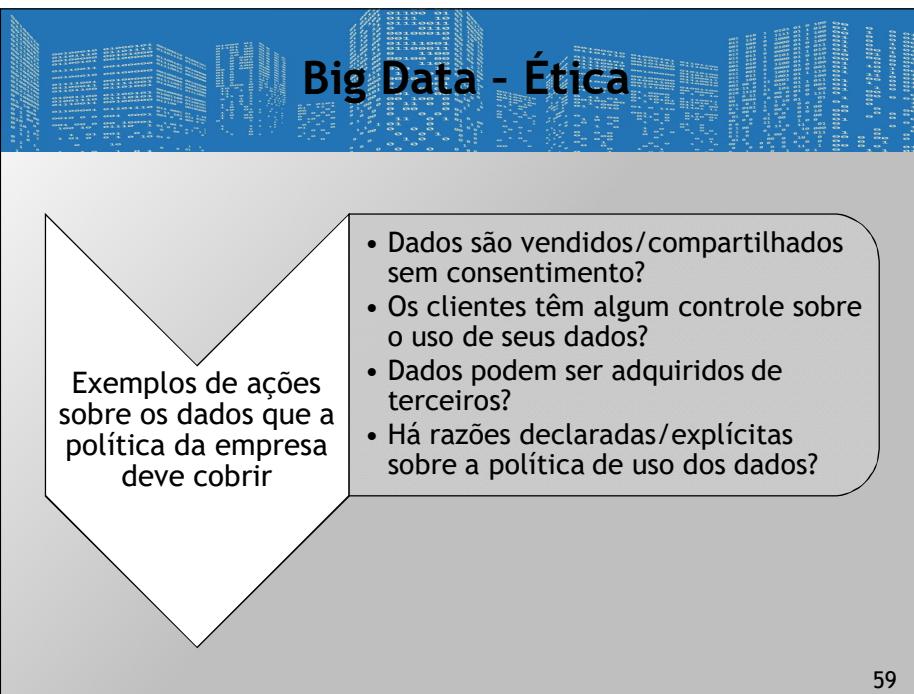
Big Data - Riscos para a Ética



Tendências do Big Data



Big Data - Ética



Bancos de Dados Pós-Relacionais - NoSQL



BD Pós-Relacionais NoSQL

Não SQL: Not Only SQL? → NoSQL

Evitar “amarração” do BD Relacional (propriedades ACID, consistências, . . .)

BASE (Basically Available, Soft-state, Eventual consistency) - Agressivo

Resposta garantida (sempre disponível)	Há estados momentaneamente inconsistentes	O estado final deve ser consistente
--	---	-------------------------------------

ACID (Atomicidade, Consistência, Isolamento, Durabilidade)

Foco na consistência - Conservador

Características NoSQL

❖ Teorema CAP

- Nem sempre a consistência dos dados é mantida
- Base no teorema CAP (Consistency, Availability e Partition tolerance) - criado por Eric Brewer
 - ✓ Consistência eventual após a execução de uma operação
 - ✓ Disponibilidade mesmo com falhas em nós
 - ✓ Capacidade de continuar operando mesmo após falha na rede
- Em um dado momento só é possível garantir **duas das três propriedades**
- Normalmente as duas últimas são privilegiadas

62

Teorema CAP

- Sistemas CA: consistência forte e alta disponibilidade, não tratam possíveis falhas de partições
- Sistemas CP: consistência forte e tolerância a falhas; abrem mão da disponibilidade
- Sistemas AP: alta disponibilidade e tolerância a falhas; abrem mão da consistência

Características NoSQL

❖ Consistência Eventual

- Propriedades ACID não são respeitadas simultaneamente
- BASE → Uma aplicação não tem de ser sempre consistente - o sistema torna-se consistente no momento devido
- O sistema deve tolerar inconsistências temporárias → priorizar a disponibilidade

❖ Quando o ACID não é necessário?

- Dados não são críticos no tempo
- Necessária rápida leitura/escrita - sem bloqueios

63

64

BD Pós-Relacionais NoSQL

❖ NoSQL ou NewSQL

➤ Chave-Valor



➤ Documentos



➤ Colunas



➤ Grafos



65

BD de Documentos

❖ Similar ao BD Chave Valor

❖ O documento corresponde ao valor

❖ O conteúdo dos documentos pode também ser usado para a recuperação dos dados armazenados

Chave	Valor
1	Doc. 1
2	Doc. 2
3	Doc. 3

67

BD Chave Valor

- ❖ O modelo de dados é uma coleção de pares chave-valor
- ❖ Chave representa um campo que vai apontar para os valores dos demais campos
- ❖ Boa escalabilidade
- ❖ Armazenamento de quaisquer valores sob uma chave
- ❖ O valor pode ser texto, blob, xml, . . .

Chave	X	Y	Z
1	A	B	C
2	R	S	T
3	M	N	O

Relacional

Chave	Valor
1	X:A; Y:B; Z:C
2	X:R; Y:S; Z:T
3	X:M; Y:N; Z:O

Chave-Valor

66

BD Orientados a Colunas

- ❖ Trabalha sobre colunas
- ❖ Cada coluna é tratada individualmente
- ❖ Valores de uma coluna são armazenados de forma contígua
- ❖ Linhas podem ser construídas a partir de colunas, se necessário
- ❖ Melhora o desempenho de consultas que visem apenas uma/poucas colunas

68

BD Orientados a Colunas

Chave	X	Y	Z
1	A	B	C
2	R	S	T
3	M	N	O

Relacional

Chave	Col1	Col2	Chave	Col1	Col2	Chave	Col1	Col2
-------	------	------	------	-------	------	------	------	-------	------	------	-------

Armazenamento

Chave	X	Chave	Y	Chave	Z
1	A	1	B	1	C
2	R	2	S	2	T
3	M	3	N	3	O

Colunar

Chave	Col1	Col1	Chave	Col2	Col2	Chave	Col3	Col3
-------	------	------	------	-------	------	------	------	-------	------	------	------

Armazenamento

69

BD de Grafos

- ❖ Armazenar entidades (nós), com suas propriedades (atributos) e seus relacionamentos (arestas direcionadas)
- ❖ Entidades têm propriedades
- ❖ Um nó é uma instância de um objeto
- ❖ Cada par de nós pode ser conectado por mais de uma aresta
- ❖ Permite, de forma livre, a adição de novas arestas e nós
- ❖ Relacionamentos: têm uma direção, um nome, um nó de origem e um de destino

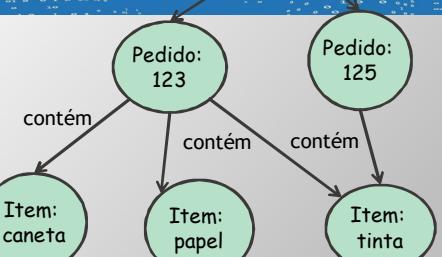
70

Usuário:
Ivo

BD de Grafos

colocou

colocou



De st.	Ori g.	Ivo	123	125	caneta	papel	tinta
Ivo	0	1	1	0	0	0	0
123	-1	0	0	1	1	1	
125	-1	0	0	0	0	1	
caneta	0	-1	0	0	0	0	
papel	0	-1	0	0	0	0	
tinta	-1	-1	0	0	0	0	

71

Hadoop
Histórico

72

Hadoop - Histórico

2003: Google cria o MapReduce - otimizar indexação e catalogação de dados (Jeffrey Dean e Sanjay Ghemawatt)

2003: publicação do artigo “The Google File System” (Sanjay Ghemawatt, Howard Gobioff, Shun-Tak Leung) - sistema de arquivos distribuído

2004: artigo “Simplified Data Processing Large Clusters” (Jeffrey Dean e Sanjay Ghemawatt)

2005: projeto Nutch - parte do projeto Lucene (comunidade de software livre) (Douglas Cutting) - versão do MapReduce

2006: Yahoo contrata Cutting → projeto Hadoop - Apache Software Foundation sob patrocínio da Yahoo

2008: Yahoo quebra recorde de processamento com o Hadoop → ordenação de 1 TB de dados em 910 máquinas, em 209 segundos

Hadoop - Histórico

❖ Histórico

- 2012: versão 1.0.1
- 2013: versão 2.0
- 2018: versão 3.0
- 2019: versão 3.1.3



❖ Quem utiliza

- Adobe: 80 nós
- E-Bay : 530 nós
- Facebook : 1.400 nós
- LinkedIn : 1.900 nós
- Yahoo : 40.000 nós
- Twitter
- New York Times

74

Hadoop HDFS e MapReduce



75

Hadoop



Hadoop: software *open source* coordenado pela *Apache Software Foundation*.

Armazenamento e processamento distribuído de grandes conjuntos de dados em clusters de computadores “comuns”

Framework para ambientes distribuídos usado principalmente para análise de grandes volumes de dados

HDFS (*Hadoop Distributed File System*) → Armazenamento de dados confiável (redundante e com alta disponibilidade)

MapReduce: → Computação paralela de alto desempenho
Processamento de conjuntos de dados, reduzindo o volume da saída de dados

Common Utilities: conjunto de bibliotecas e utilitários comuns que auxiliam outros módulos Hadoop

Hadoop - 3 Serviços Principais

Map Reduce
(Computação Distribuída)



HDFS
(Armazenamento Distribuído)



Common Utilities

77

Hadoop - Características

Código aberto: comunidade ativa, apoio de grandes empresas, evolução contínua

Econômica: software livre, máquinas convencionais, serviços na nuvem

Robustez: tem estratégias de recuperação em caso de falhas (replicação de dados)

Escalabilidade horizontal simplificada (novas máquinas)

Tolerante a falhas: a perda de um nó não paralisa o processamento

78

Hadoop



- ❖ Roda em grande número de máquinas, sem compartilhamento de memória ou de discos
- ❖ Mantém a trilha do armazenamento dos arquivos
- ❖ **HDFS - Hadoop Distributed File System**
 - Suporta qualquer tipo de dados
 - Gerencia o conjunto de servidores
 - Armazenamento de grandes conjuntos de dados de forma distribuída
- ❖ **Map-Reduce**
 - Dados são divididos em porções
 - Porções são processadas em paralelo
 - Geram resultados intermediários
 - Resultado final agrupa tais resultados intermediários

79

Hadoop



Web logs

Imagens/Vídeos

Mídias Sociais

Documentos



Big Data
NoSQL



Big Data
Warehouse



OLAP

Análise

Data Mining

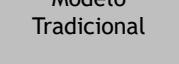
Relatórios



Data
Mining



Relatórios



Modelo
Tradicional

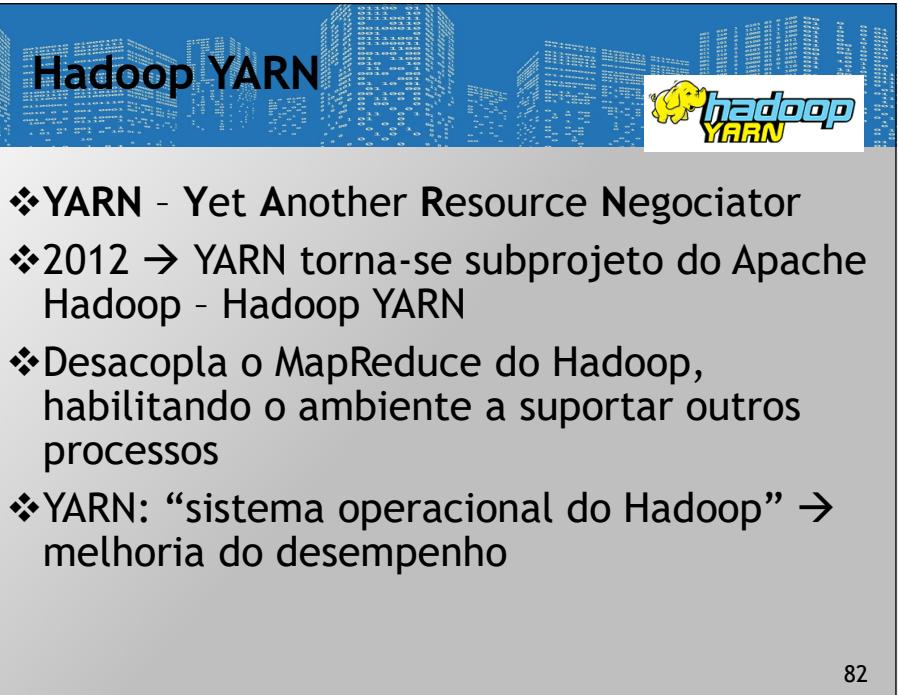
80



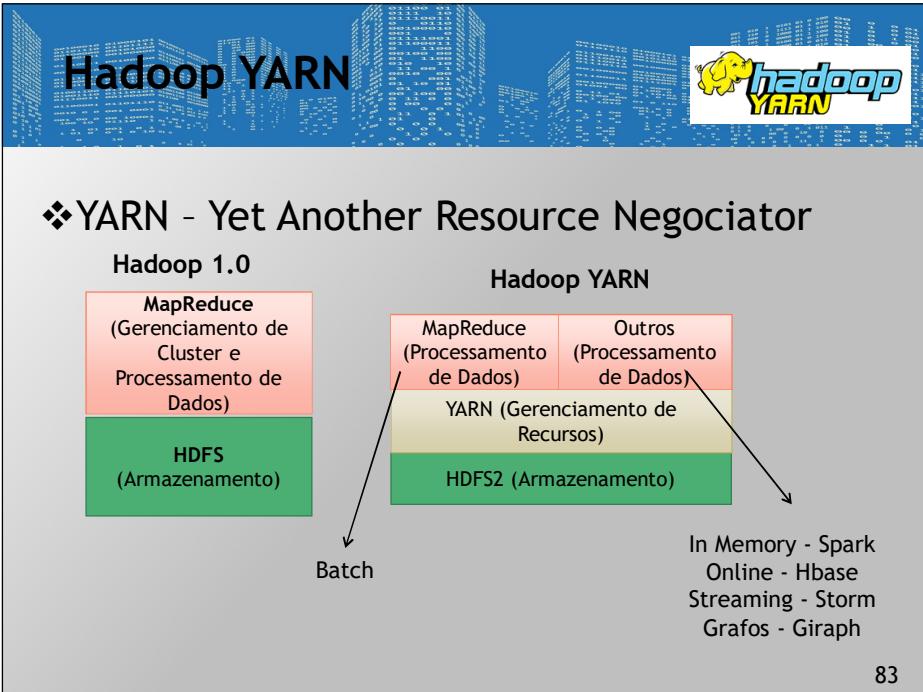
Hadoop YARN



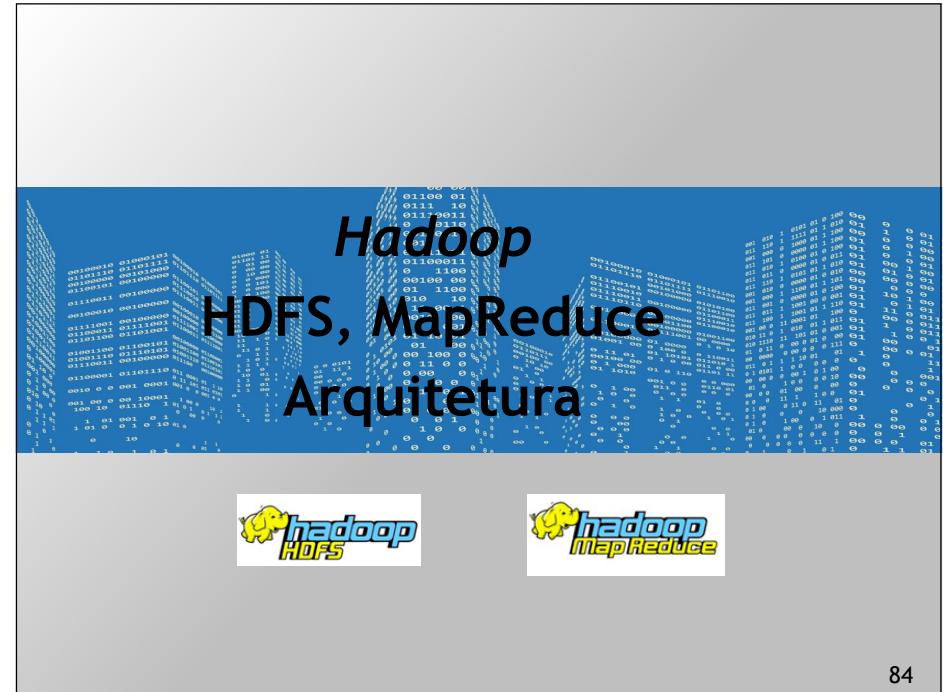
81



82

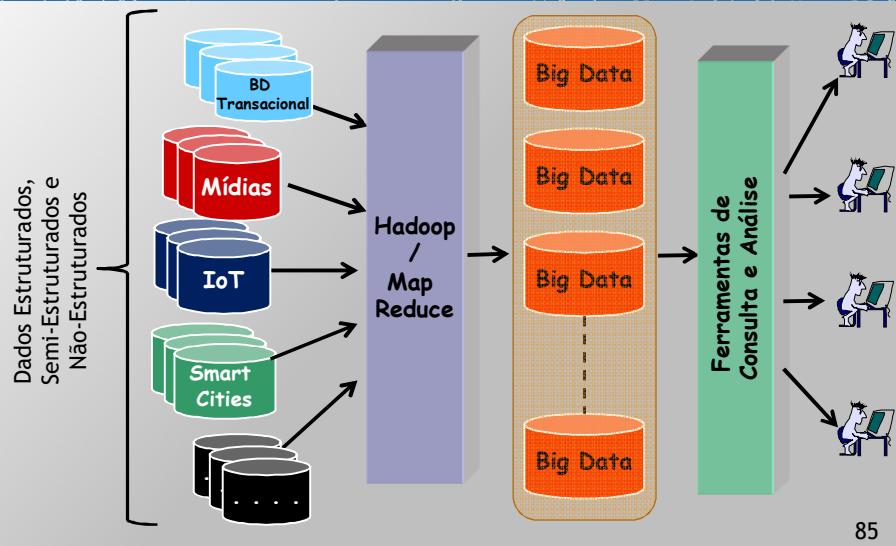


83

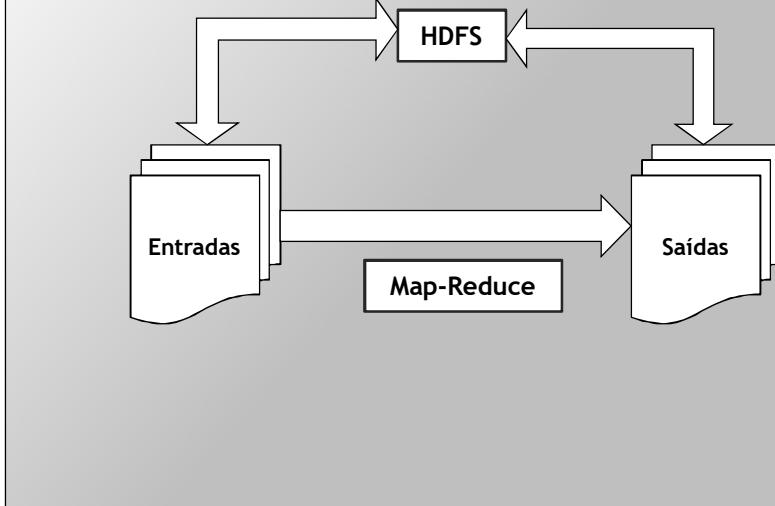


84

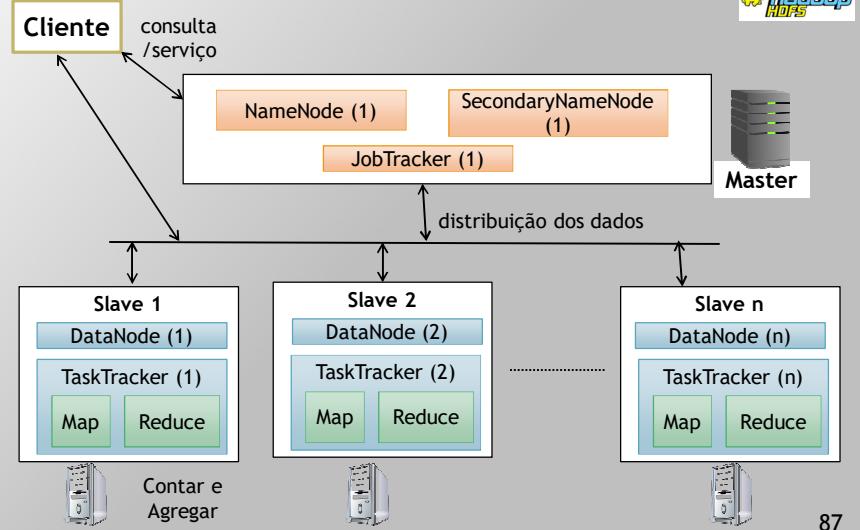
Arquitetura Big Data - Hadoop



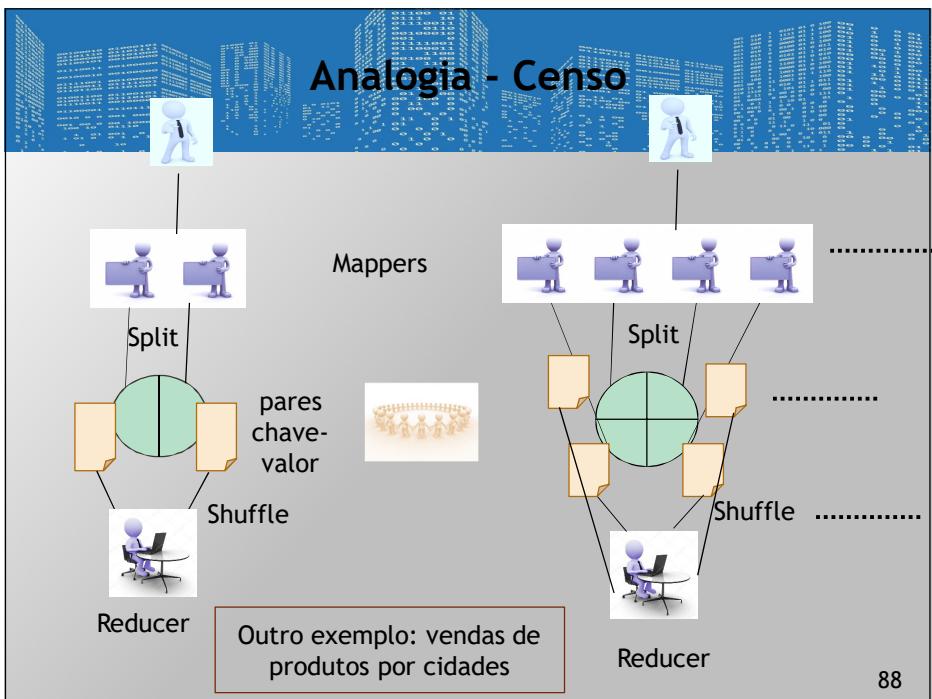
Processos Básicos do Hadoop



Processos Básicos do Hadoop



Analogia - Censo



Tarefas Executadas - Cliente

Escrever um arquivo no HDFS

Comunica-se com o NameNode para obter os metadados

NameNode responde com o número de blocos e sua localização

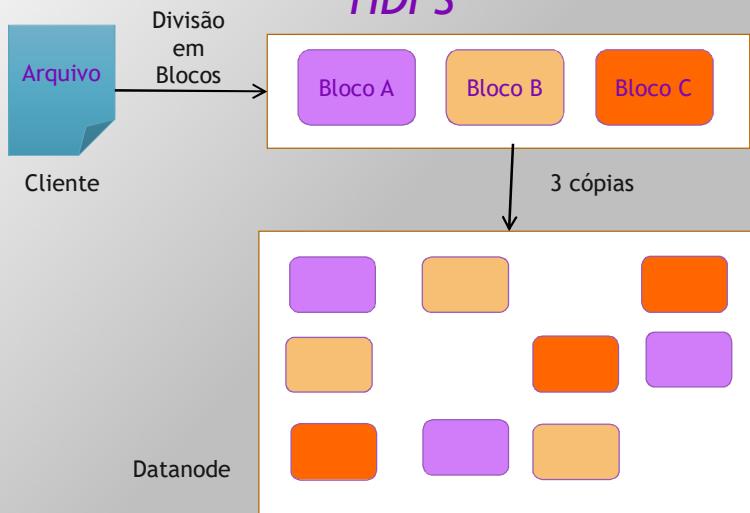
O Cliente divide o arquivo em blocos e os envia ao primeiro DataNode

Esse DataNode replica o bloco a mais 2 DataNodes e envia uma confirmação ao NameNode

89

Hadoop - MapReduce

HDFS



90

Tarefas Executadas - Cliente

Solicitar uma tarefa ao Job Tracker

Job Tracker distribui a tarefa aos Task Trackers - Fase Map

Task Trackers executam a Fase Map

Job Tracker distribui a tarefa aos Task Trackers - Fase Reduce

Task Trackers executam a Fase Reduce

91

Tarefas Executadas - Cliente

Ler um arquivo do HDFS

Interage com o NameNode para obter a localização dos blocos

Interage com os DataNodes indicados, havendo troca de mensagens e verificação de autorização

92

Tarefas Executadas - Master



- ❖ Recebe um serviço solicitado pelo Cliente
- ❖ Cria um plano de execução: determina quais Slaves serão acionados
- ❖ Gerencia a execução das tarefas (conhece os recursos consumidos e os disponíveis)
- ❖ Gerencia arquivos armazenados e sua divisão em blocos (arquivos de 128 MB)
- ❖ Controla réplicas dos dados (3 réplicas de cada arquivo)

93

Tarefas Executadas - Master



- ❖ Verifica o processamento em intervalos pré-definidos
- ❖ Modifica as atribuições em caso de falha
- ❖ Namenode armazena os metadados necessários: nome e tamanho dos arquivos, localização dos blocos, datas de criação e modificação, ...
- ❖ Novos dados/arquivos são frequentes → Namenode guarda as informações sobre a localização dos arquivos em memória

94

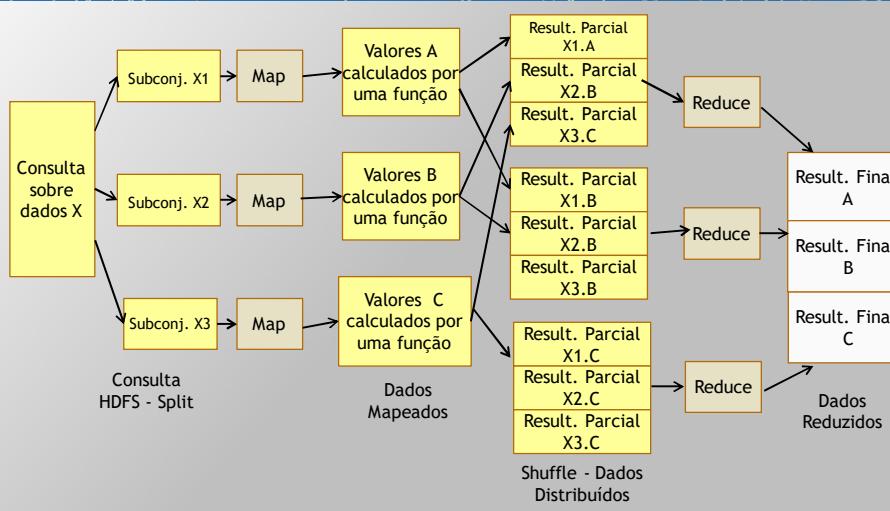
Tarefas Executadas - Slaves



- ❖ Armazenam os blocos de dados originais
- ❖ Processam grandes volumes de dados de forma paralela, dividindo o trabalho em frações independentes
- ❖ Escrita em várias linguagens: Python, C++, Java, ...
- ❖ Executam as tarefas Map e/ou Reduce - modelo de programação, não uma linguagem
- ❖ Se a comunicação com o Master atingir um limite de tempo sem resposta → assume-se falha

95

Hadoop - MapReduce



96

Modos de Execução do Hadoop



❖ Modo Local

- Configuração padrão
- Útil para testes iniciais do ambiente

❖ Modo Pseudo Distribuído

- Cluster de apenas uma máquina
- Ainda utilizado na fase de testes do ambiente

❖ Modo Completamente Distribuído

- Processamento real final

97

Apache Spark



Apache Spark



- ❖ Projeto da Apache Software Foundation
- ❖ Versão inicial: 2014
- ❖ É um framework, não um produto isolado
- ❖ APIs em Java, Scala e Python
- ❖ Players principais: Amazon, eBay, Yahoo

99

Apache Spark



❖ Map Reduce

- Consumo de tempo pela replicação, serialização e disk I/O
- Aplicações Hadoop consomem 90% do tempo realizando operações de escrita / leitura

❖ Spark

- Processamento com dados diretamente in-memory (principal diferença)
- Resultado: 10 a 100 vezes mais rápido
- Porém o volume de dados processado é menor

100

Hadoop Fatos e Mitos

101

Mitos sobre o Hadoop

Mito: Hadoop consiste de um produto simples (single)

Fato: Hadoop consiste de vários produtos - abrange uma família de produtos open source, como HDFS, MapReduce, Pig, Hive, Hbase, Hcatalog, Mahout, . . .

Mito: Hadoop tem a preocupação exclusiva com grandes volumes de dados

Fato: Hadoop também tem como foco a diversidade de dados - pode gerenciar o acesso a qualquer tipo de dados, armazenando-os no HDFS

Mitos sobre o Hadoop

Mito: Todos componentes do Hadoop são exclusivamente open source

Fato: Há também distribuições comerciais, como IBM, Cloudera e EMC - geralmente incluem ferramentas não oferecidas pela Apache, como softwares administrativos

Mito: A única ferramenta para tratar big data é o Hadoop

Fato: Não necessariamente precisa ser usado o Hadoop - há outras plataformas: HP Vertica, Teradata Unified Architecture

103

Mitos sobre o Hadoop

Mito: o HDFS é o gerenciador de dados do Hadoop

Fato: O HDFS é um sistema de arquivos e não um gerenciador de bases de dados - podem ser usados como gerenciadores o Hbase ou o Hive

Mito: O MapReduce faz parte do HDFS

Fato: HDFS e MapReduce fazem uma boa parceria, mas um pode ser usado sem o outro

Mito: Hadoop é uma alternativa/substituto do Data Warehouse

Fato: Hadoop pode auxiliar um Data Warehouse, mas não substituir

104

A slide comparing Data Warehouse and Big Data. The title is "Comparação Data Warehouse x Big Data". The background features a blue city skyline where each building is composed of binary code. The slide is numbered 105 at the bottom right.

DW x Big Data	
DW	Big Data
Arquitetura para armazenamento e consulta de dados	Tecnologia para tratar e consultar grandes conjuntos de dados
Processamento centralizado	Processamento distribuído
Análise de questões importantes das atividades da empresa / organização	
Dados estruturados	Dados estruturados e semi-estruturados
Bases de dados relacionais	Bases de dados relacionais e NoSQL

105

106

A slide featuring the word "Data Lake" in large, bold, black letters. The background is a blue city skyline with binary code on the buildings. The slide is numbered 107 at the bottom right.

A slide titled "Definição Original - James Dixon". It includes a quote from James Dixon's blog: "If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples." It also shows a timestamp: "Written by James October 14, 2010 at 4:06 pm". The slide is numbered 108 at the bottom right.

107

108

Definição

❖ Data Lake é um repositório enorme que acolhe todos os tipos de dados em seus formatos nativos, sendo que, em algum momento será utilizado por alguém da organização

109

Esquema On Read e On Write

A estrutura dos dados deve ser totalmente definida antes da escrita de qualquer dado: tabelas, colunas, chaves, ...

Inclui tipos de dados e tamanhos

Estrutura otimizada para consultas mais rápidas

Dificuldade de se implementar mudanças nessa estrutura

111

Não é necessário definir a priori a estrutura dos dados

O esquema é definido quando os dados são lidos, mas não quando são armazenados

No momento da leitura define-se quais serão os dados necessários a cada objetivo

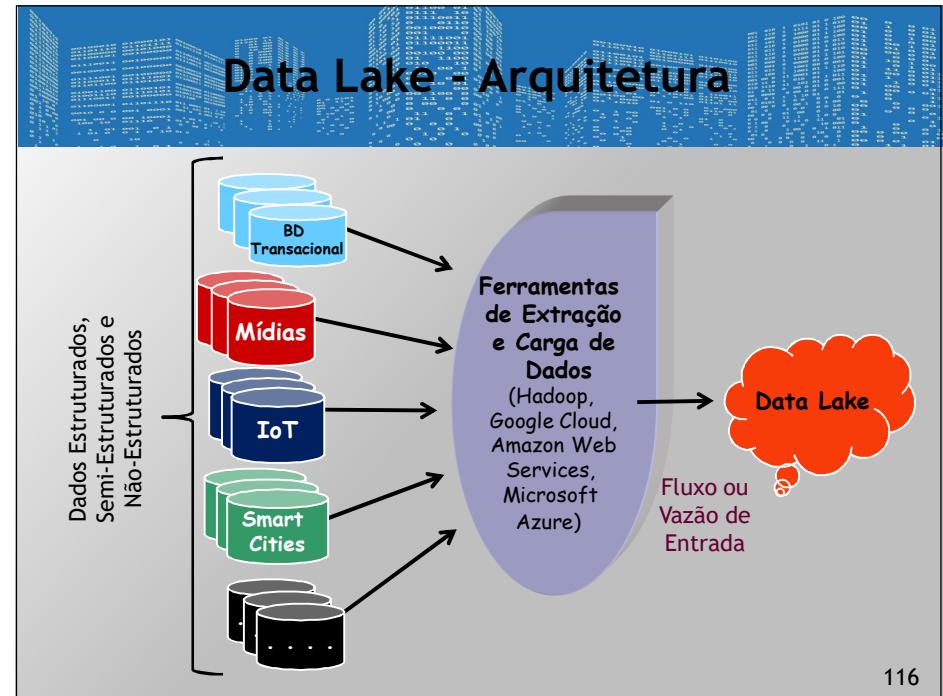
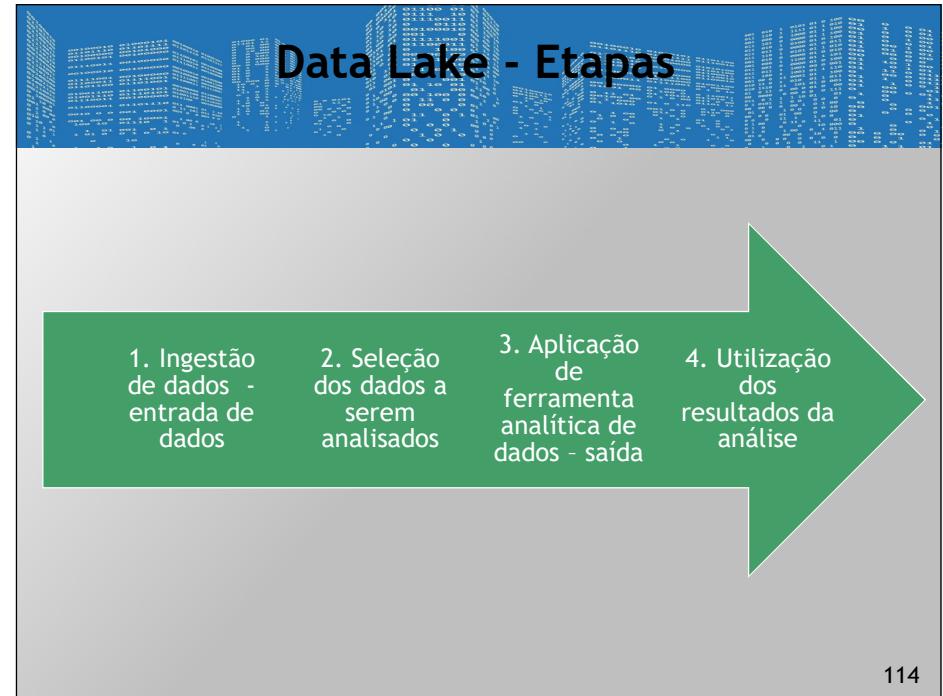
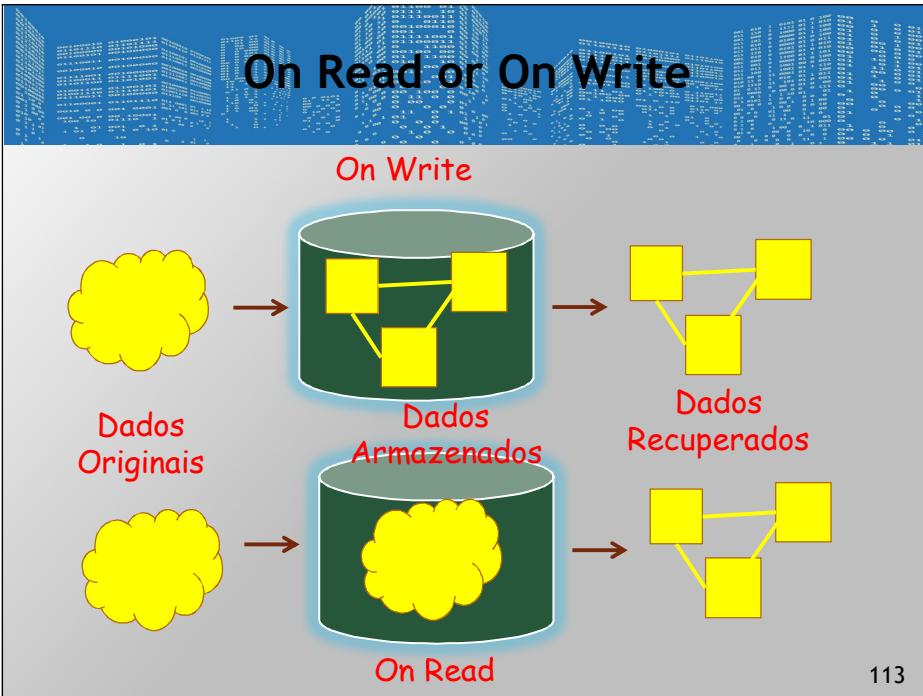
Dados não são alterados ou descartados no momento do armazenamento

Sem o ETL, os dados devem ser tratados no momento da leitura → consultas mais lentas

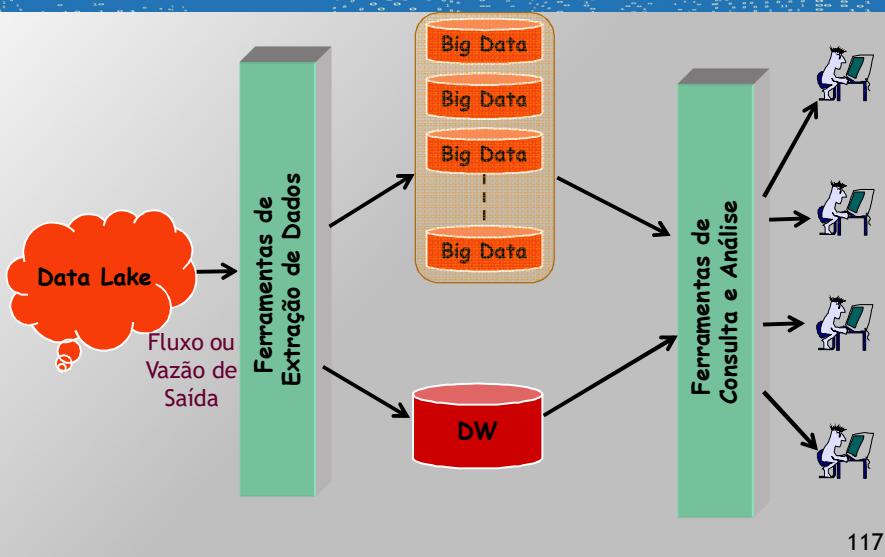
110

Esquema On Read

112

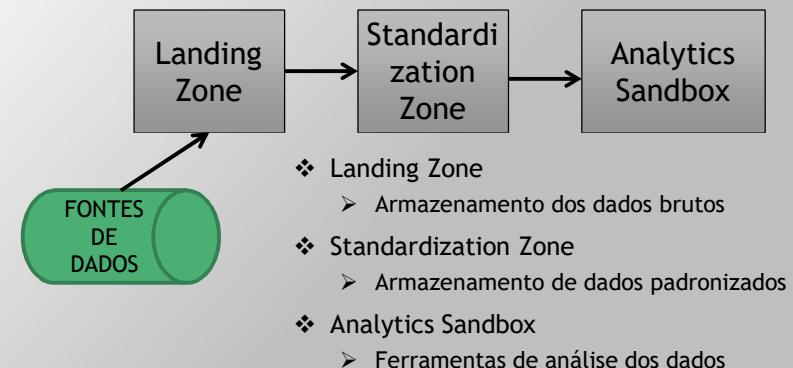


Data Lake - Arquitetura



117

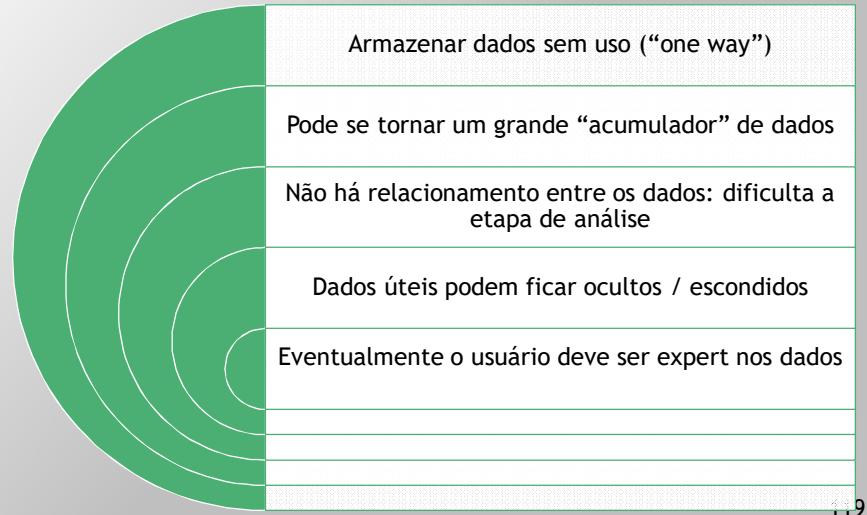
Data Lake - Arquitetura Outra Nomenclatura



OU: Landing Zone, Gold Zone e Work Zone

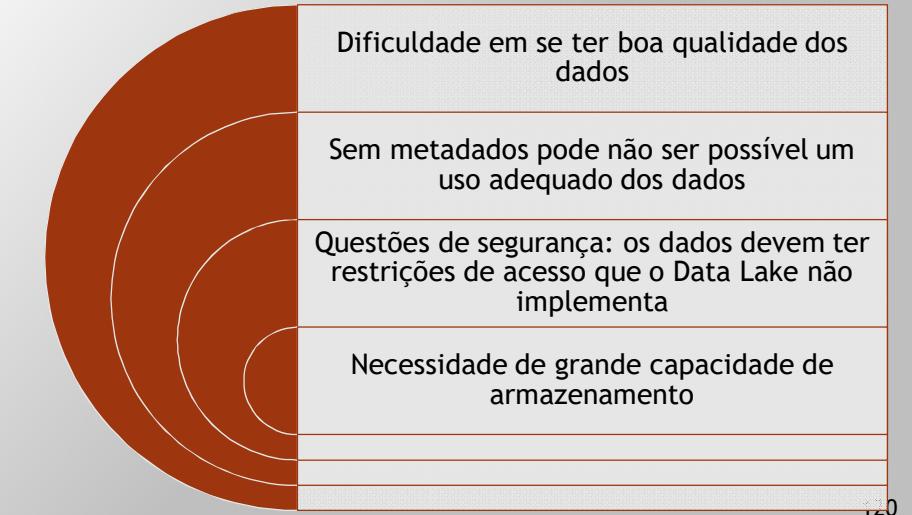
118

Riscos do Data Lake



119

Riscos do Data Lake



120

Data Lake - Resumo

Armazenar dados sem se preocupar como vai ser seu uso: “Ingestão” de dados em qualquer formato

Dados estruturados, semiestruturados e não estruturados, todos armazenados na forma “nativa”

Armazenar também os metadados

Servir como um ODS (Operational Data Store)

Na análise dos dados → necessário definir / adotar algum esquema

Organização dos dados: → momento da análise

Tipos de dados: Web, Sensores, Logs, Redes Sociais, Imagens, ...

121

Data Lake - Metadados

- ❖ Metadados: etiquetar todos os dados brutos
- ❖ Tem um mapeamento dos dados
- ❖ Conhecer o contexto/origem dos dados
- ❖ Ter, a priori, uma indicação de como tratar os dados

122

Usuários

Cientistas de Dados

- Data Lake
- Big Data

Analistas de Negócio

- Data Warehouse
- Big Data

123

Transformar o Data Lake na “Mina de Ouro”

- ❖ Hadoop
- ❖ Amazon Web Services (AWS)
- ❖ Microsoft Azure
- ❖ Google Cloud Platform

124

Casos de Uso - Óleo e Gás

- ❖ Uma das primeiras a adotar data lakes na nuvem
- ❖ Gera 1.5 TB de dados IoT por dia
- ❖ Objetivo: aumento da produção
 - Otimização de perfurações
 - Minimização de tempos de parada não planejados
 - Segurança aumentada
 - Aderência a regulações

125

Casos de Uso - Smart Cities

- ❖ Objetivos
 - Melhora do tráfego
 - Orientar a aplicação de leis
 - Otimizar sistemas de educação
 - Otimizar redes elétricas, hidrovias, pedágios, . . .
- ❖ Exemplo: um veículo conectado gera 25 GB de dados por hora na nuvem

126

Casos de Uso - Ciências da Vida

- ❖ Possibilidade de aumentar a expectativa de vida
- ❖ Genoma humano: 175 a 225 GB por pessoa
- ❖ Outros dados: frequência cardíaca, pressão sanguínea, contage de glóbulos brancos, temperatura corporal, . . .
- ❖ Exames de imagens: tomografias, ressonâncias, ultrassom, densitometria,

127

“Data Lakehouse”

- ❖ Meio caminho entre Data Warehouse e Data Lake
- ❖ Utilizar dados estruturados e semi-estruturados
- ❖ Usuários: cientista de dados e analistas de dados
- ❖ Flexibilidade: Schema on Read e Schema on Write
- ❖ Tese de Pedro Javier Gonzales Alonso, em 2016, Barcelona

128

Comparação Data Warehouse x Data Lake

129

DW x Data Lake / Big Data	
DW	Data Lake
Arquitetura para armazenamento e consulta de dados	Tecnologia para tratar e consultar grandes conjuntos de dados
Processamento centralizado	Processamento distribuído
Análise de questões importantes das atividades da empresa / organização	
Dados estruturados	Dados estruturados, semi-estruturados e não estruturados
Esquema de dados de difícil alteração	Sem esquema
Consistência alta	Consistência baixa

DW x Data Lake / Big Data

DW	Data Lake
Usuários: business	Usuários: cientistas de dados
Esquema on write: processa os dados antes de armazenar → “dados limpos”	Esquema on read: armazena os dados antes de processar → “dados brutos”
Baixa granularidade	Alta granularidade
Expectativa de uso de todos os dados	Expectativa de uso de parte dos dados

131

Big Data e Data Lake Resumo Arquitetura

132

Arquitetura Big Data e Data Lake

