

## Frameworks de Big Data

**Eduardo Viegas**



1

## Eduardo Viegas

- **Formação**
  - Graduado em Ciência da Computação – PUCPR
  - Mestre em Informática – PUCPR
  - Doutorado em Informática – PUCPR
    - Universidade de Lisboa (Distributed Research Team)
- **Membro da Intel Strategic Research Alliance (ISRA) Brasil**
  - Energy-efficient Security for SoC Devices
- **Atuação**
  - Computação em nuvem
  - Big Data
  - Segurança da informação
  - Aprendizagem de máquina
  - Pesquisa na indústria e academia
- **Professor pesquisador no PPGLa PUCPR**

2/12/2020 Machine Learning – Eduardo Viegas 2

2

## Cronograma das aulas

Data	Conteúdo
01/02	Big Data, Hadoop e HDFS
15/02	HDFS, MapReduce
29/02	HDFS, MapReduce, Apache Spark
14/03	HDFS, Apache Spark
21/03	HDFS, Apache Spark, Projeto

Big Data – Eduardo Viegas 3

3

## Avaliação

- **30% da nota**
  - 2 atividades em sala
- **70% da nota**
  - Trabalho em equipe
  - Resolução de problema prático envolvendo Big Data e tecnologias estudadas
  - Relatório em formato SBC de 10 páginas

Big Data – Eduardo Viegas 4

4

## Antes de começarmos =)

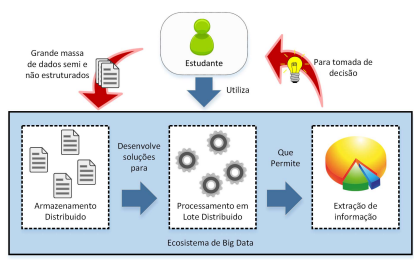
- **Quem aqui sabe programar?**
  - **Python?**
  - **BASH?**

Big Data – Eduardo Viegas 5

5

## Antes de começarmos =)

- **O que faremos nesta disciplina?**



Big Data – Eduardo Viegas 6

6

## Antes de começarmos =)

### Ao final desta disciplina você deverá saber

- Reconhecer cenários de aplicação de Big Data, considerando dados semi e não estruturados.
- Reconhecer os principais componentes dos ambientes de HDFS e HADOOP e suas funcionalidades.
- Aplicar técnicas de processamento em Lote para análise de dados em Big Data.
- Aplicar técnicas de estruturação e análise de dados estruturados ou semi estruturados
- Desenvolver técnicas de Big Data para análise de dados

Big Data – Eduardo Viegas

7

7

## Agenda

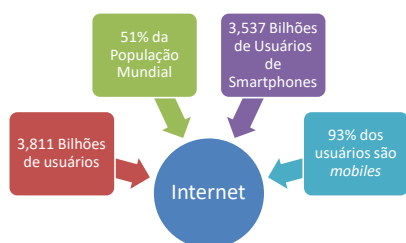
- **Big Data**
- Hadoop
  - HDFS
  - MapReduce
  - Apache Spark

Big Data – Eduardo Viegas

8

8

## Massa de dados ao longo do tempo



Big Data – Eduardo Viegas

9

9

## Massa de dados ao longo do tempo

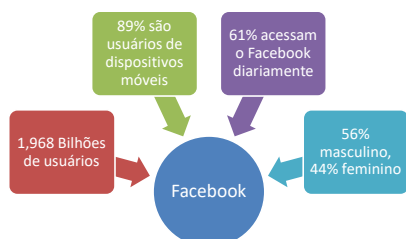


Big Data – Eduardo Viegas

10

10

## Massa de dados ao longo do tempo

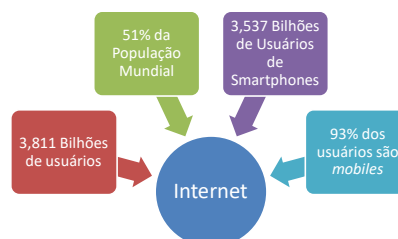


Big Data – Eduardo Viegas

11

11

## Massa de dados ao longo do tempo



Big Data – Eduardo Viegas

12

12

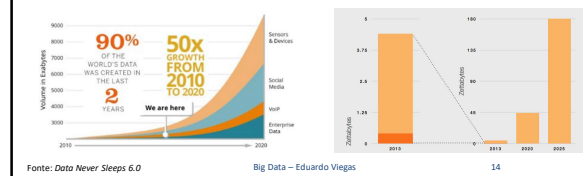
## A cada minuto



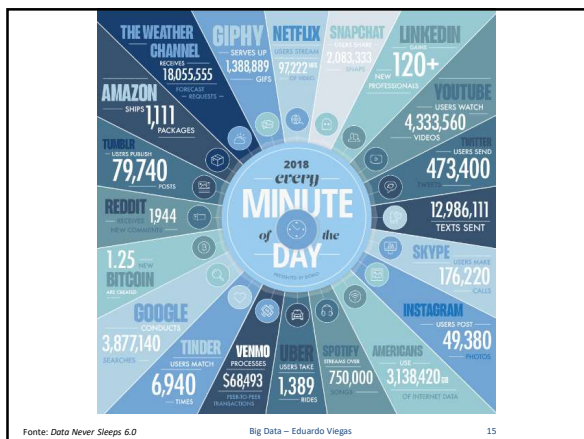
13

## Volume

- 90% dos dados foram gerados nos últimos 2 anos
  - Se manteve assim por 30 anos
- Espera-se que em 2025, os dados produzidos em uma semana serão equivalentes aos dados produzidos pela humanidade até 2013
  - Smartphones
- Criamos muitos dados sem saber
  - Posts, tweets, fotos, vídeos, ...
  - Logs, sensores, ...



14



15



16

## Big Data

- "...Um conjunto de dados tão grande e complexo que torna o seu processamento e armazenamento através de técnicas tradicionais impraticável..." *NIST*
- "...Big Data é grande Volume, alta Velocidade, e alta Variedade de ativos de informação que exigem formas inovadoras de baixo custo de processamento de informação para melhor percepção e tomada de decisão..." *Gartner*
- "... Big Data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it..." *Dan Ariely*
- "... Quando o Excel trava abrindo o arquivo..." *Aluno de 3º período*

Big Data – Eduardo Viegas

17

17

## Big Data

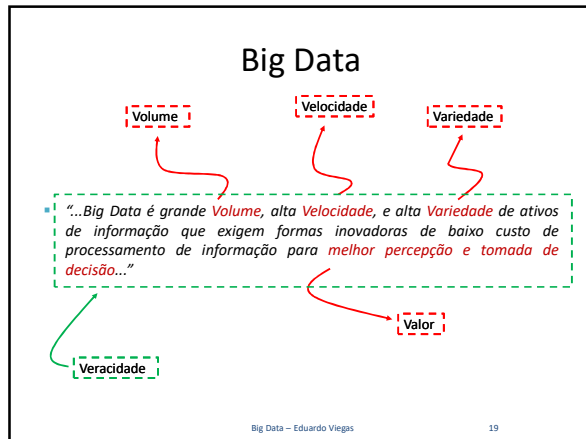
- "...Um conjunto de dados tão grande e complexo que torna o seu processamento e armazenamento através de técnicas tradicionais impraticável..." *NIST*



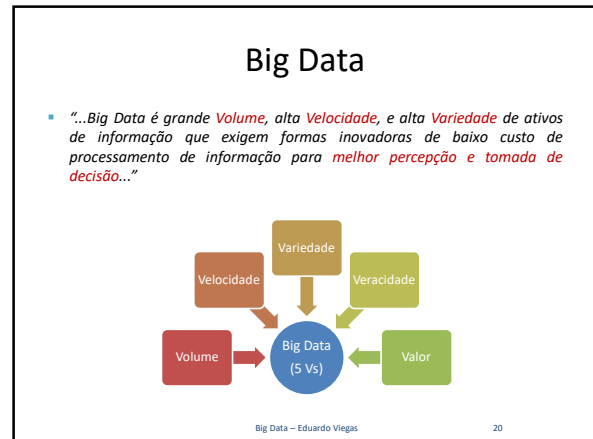
Big Data – Eduardo Viegas

18

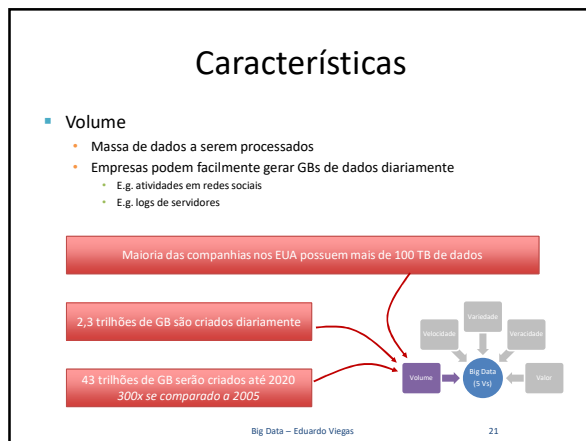
18



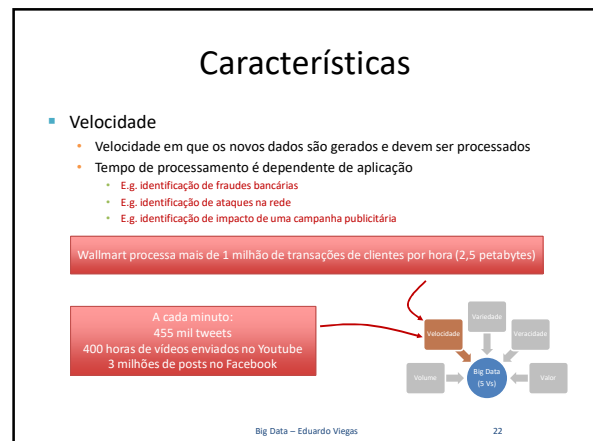
19



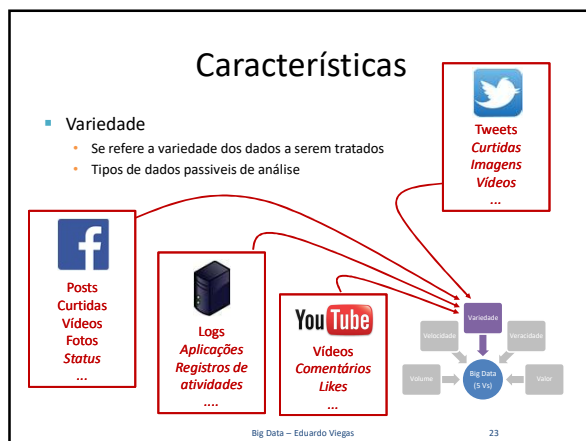
20



21



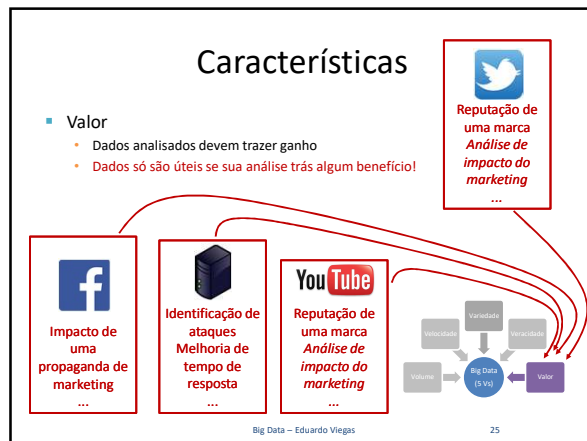
22



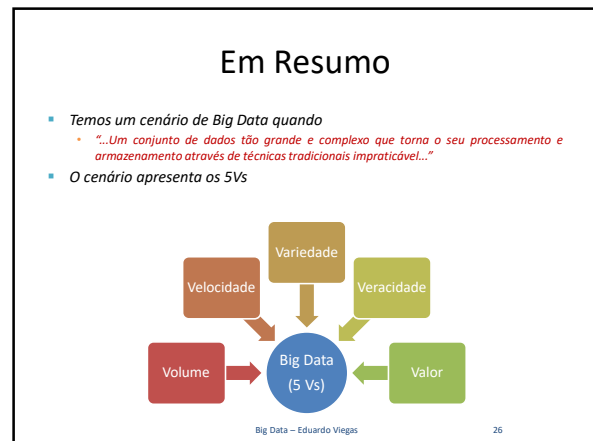
23



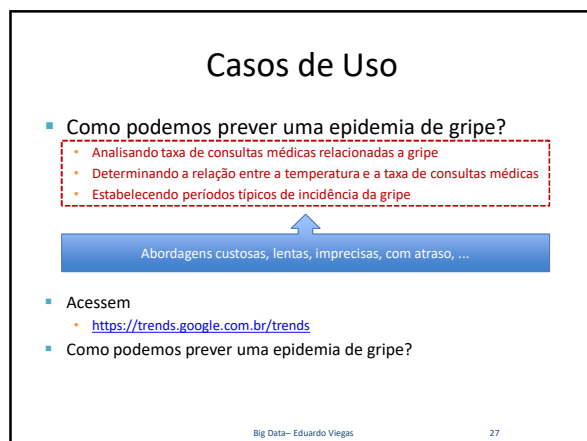
24



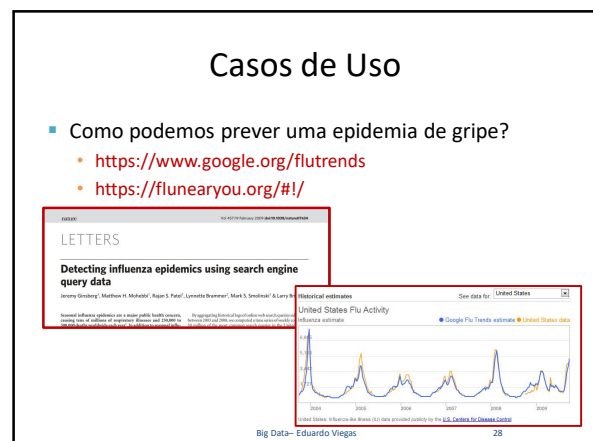
25



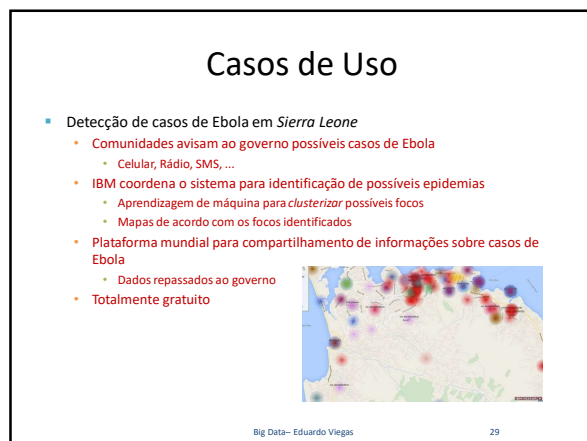
26



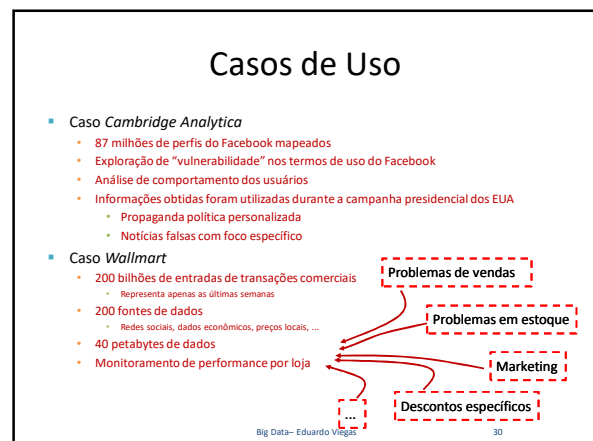
27



28



29



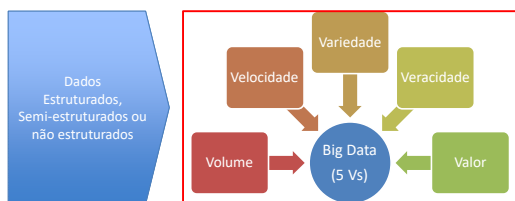
30





## Em Resumo

- Temos um cenário de Big Data quando
  - "...Um conjunto de dados tão grande e complexo que torna o seu processamento e armazenamento através de técnicas tradicionais impraticável..."
- O cenário apresenta os 5Vs



Big Data – Eduardo Viegas

43

43

## Agenda

- Big Data
- **Hadoop**
  - HDFS
  - MapReduce
  - Apache Spark

kahoot.it

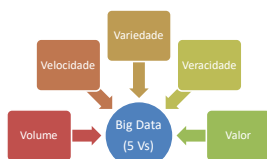
Big Data – Eduardo Viegas

44

44

## Infraestrutura

- Como processar dados com as características de Big Data?



- As características dos dados de Big Data implicam em novos desafios de infraestrutura

Big Data – Eduardo Viegas

45

45

## Infraestrutura

- Torna-se necessário uma infraestrutura distribuída de processamento e armazenamento
- Principais Desafios
  - Processamento lento, falta de escalabilidade
  - Busca em disco para cada leitura/escrita
  - Velocidade de leitura/escrita em disco se torna um gargalo
  - Armazenamento e processamento de grandes massas de dados



IDE – 75MB/s

SATA – 300MB/s

SSD – 800MB/s

- Análise, processamento, agregação, atraso de processamento, ...

Big Data – Eduardo Viegas

46

46

## Infraestrutura

- Torna-se necessário uma infraestrutura distribuída de processamento e armazenamento
- Principais Desafios
  - Falta de confiabilidade
    - 1000 máquinas processando/armazenando os dados, 1000 máquinas podem falhar
    - Restauração de dados
    - Restauração de nós
  - Escalabilidade
  - Backup
  - Custo
  - Facilidade de uso
  - Facilidade de processamento
  - Processamento distribuído
  - Armazenamento distribuído

Big Data – Eduardo Viegas

47

47

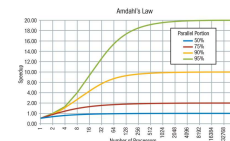
## Infraestrutura

- Processamento de dados em paralelo?



- Paralelismo não é tão simples!

- Coordenação
- Deadlock
- Sincronização
- Capacidade de rede limitada
- Divisão e agregação
- Disponibilidade



Big Data – Eduardo Viegas

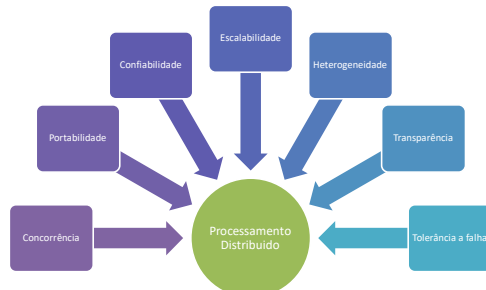
48

48



## Infraestrutura

- Processamento de dados através de computação distribuída?



Big Data – Eduardo Viegas

49

49



- Apache Hadoop é um framework que permite o processamento distribuído de grandes massas de dados, utilizando um *cluster* convencional através de um modelo simples de programação.
- Foi desenvolvido para permitir **processamento e armazenamento escalável** utilizando milhares de máquinas
- Hadoop torna o uso de uma **infraestrutura distribuída** transparente ao usuário

Eficiente
Confiável
"fácil" uso
Código aberto
Mantido como um projeto Apache
Suportado por grandes companhias

Big Data – Eduardo Viegas

50

50



- Sistema de arquivo tolerante a falha
  - Hadoop Distributed File System (HDFS)
  - Inspiração no Google File System
- Se inspira no modelo baseado em localidade
  - Leva o processamento aos dados
  - Diminui uso de rede
  - Diminui acesso a disco
- Escalabilidade
  - Programa executado é o mesmo para 1, 10, 100, 1000, ... nós
  - Performance escalável
- Modelo de computação MapReduce

Técnica precursora

Big Data – Eduardo Viegas

51

51



- Princípios do Hadoop

Falhas devem ser tratadas transparentemente ao usuário

- Sistema deve se gerenciar e se recuperar

Tarefas executadas de acordo com a performance dos nós

- Performance deve ser linearmente escalável

Proporcional a quantidade de nós utilizados

- Processamento deve ser próximo aos dados

Princípio da localidade, redução da latência e do uso da rede

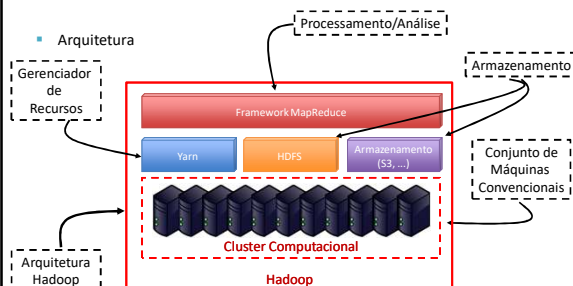
- Simples, modular e extensivo

52

52



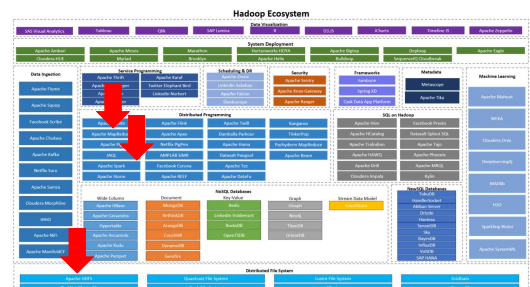
- Arquitetura



Big Data – Eduardo Viegas

53

53



Big Data – Eduardo Viegas

54

54

**APACHE hadoop**

- Hadoop possui uma arquitetura Mestre-Escravo
  - Um nó mestre gerencia um conjunto de nós escravos
  - Nó mestre é responsável pelo gerenciamento
  - Nós escravos efetuam o armazenamento/processamento

Processamento paralelo  
Escalável  
Execução de tarefas de maneira independente  
Gerenciamento de replicação  
Tolerância a falha  
Consistência de dados  
...

Mestre

Escravo

Big Data – Eduardo Viegas 55

55

**APACHE hadoop**

- Funcionalidades
  - Armazenamento de dados
    - Estruturado, semi estruturado ou não estruturado
  - Capacidade de armazenamento
    - Escala linearmente
  - Tolerância a falha
    - Confiabilidade
  - Redução de custos e complexidade
    - Utiliza computadores tradicionais
  - Complexidade é transparente ao usuário

Big Data – Eduardo Viegas 56

56

**APACHE hadoop**

- Arquitetura
  - Armazenamento no Hadoop através do HDFS

Framework MapReduce

Yarn HDFS Armazenamento (S3, ...)

Cluster Computacional

Hadoop

Big Data – Eduardo Viegas 57

57

## Agenda

- Big Data
- Hadoop
  - HDFS**
  - MapReduce
  - Apache Spark

Big Data – Eduardo Viegas 58

58

## Armazenamento em Big Data

- Como vocês armazenariam 1 PB de dados?
  - 1000 Terabytes
  - Composto por um único arquivo

x1000 ...

Big Data – Eduardo Viegas 59

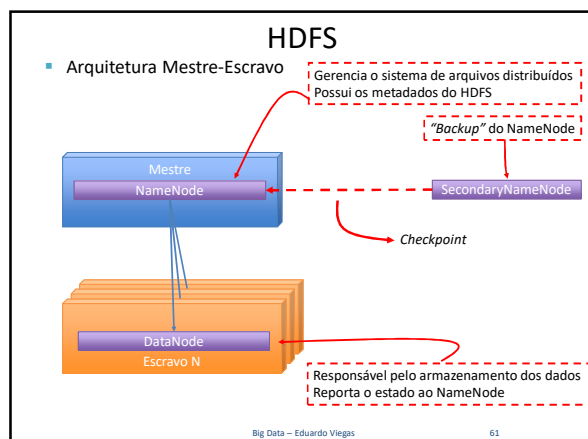
59

## HDFS

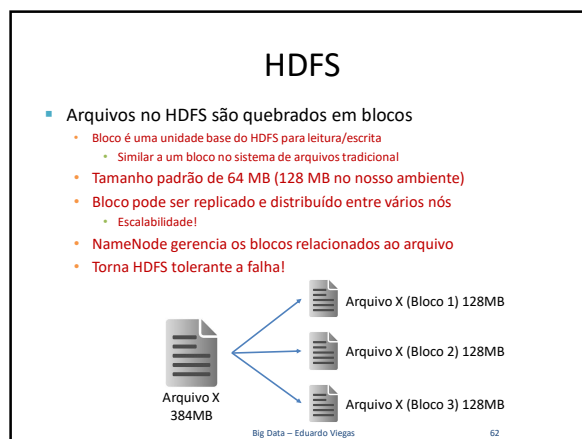
- Hadoop Distributed File System (HDFS)
- Ideal para
  - Grandes massas de dados
  - Acesso de maneira paralela aos dados
- Não recomendável para
  - Diversos arquivos pequenos
  - Acessos aleatórios aos arquivos
  - Leitura de baixa latência

Big Data – Eduardo Viegas 60

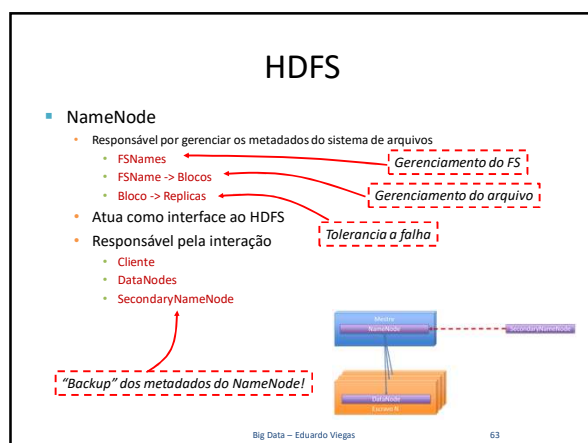
60



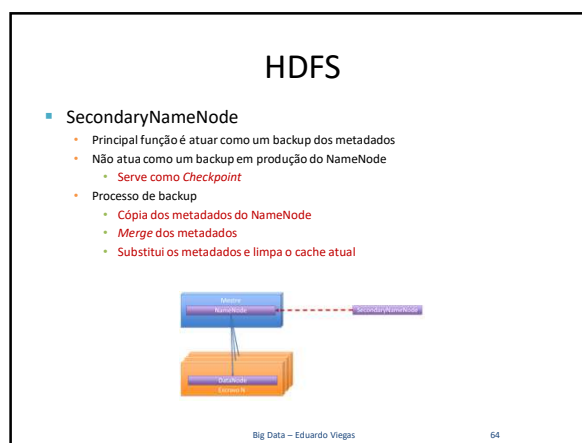
61



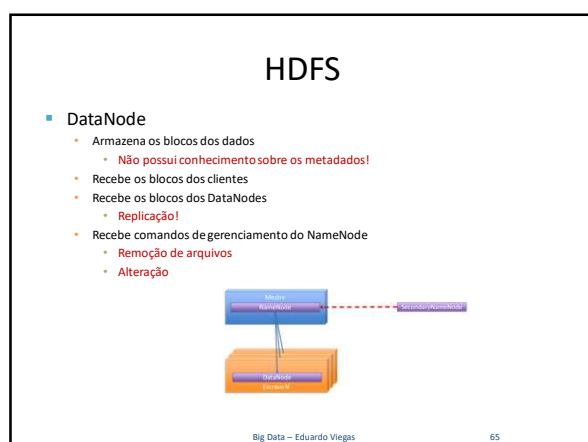
62



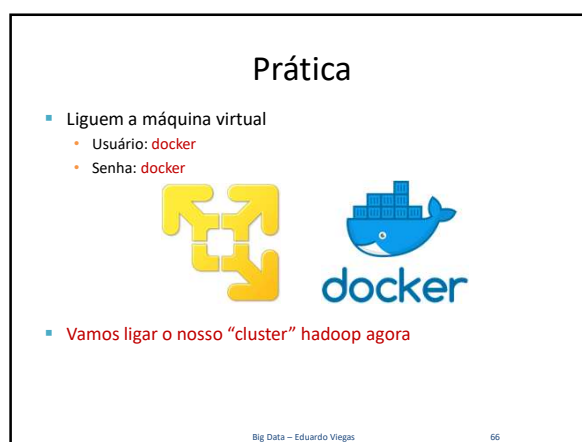
63



64



65

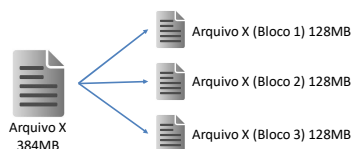


66

## HDFS

### ■ Prática HDFS

- Acesse o nosso servidor
- Crie um arquivo no sistema de arquivos local
  - `fallocate -l 1G teste.img`
- Crie uma pasta no HDFS
- Copie o arquivo para o HDFS
- Acesse cada datanode, e note os blocos em `/hadoop/dfs/data/current`
- Recupere o arquivo, exclua do HDFS, e note novamente os blocos
- Encerre a execução de um datanode



67