BIG DATA

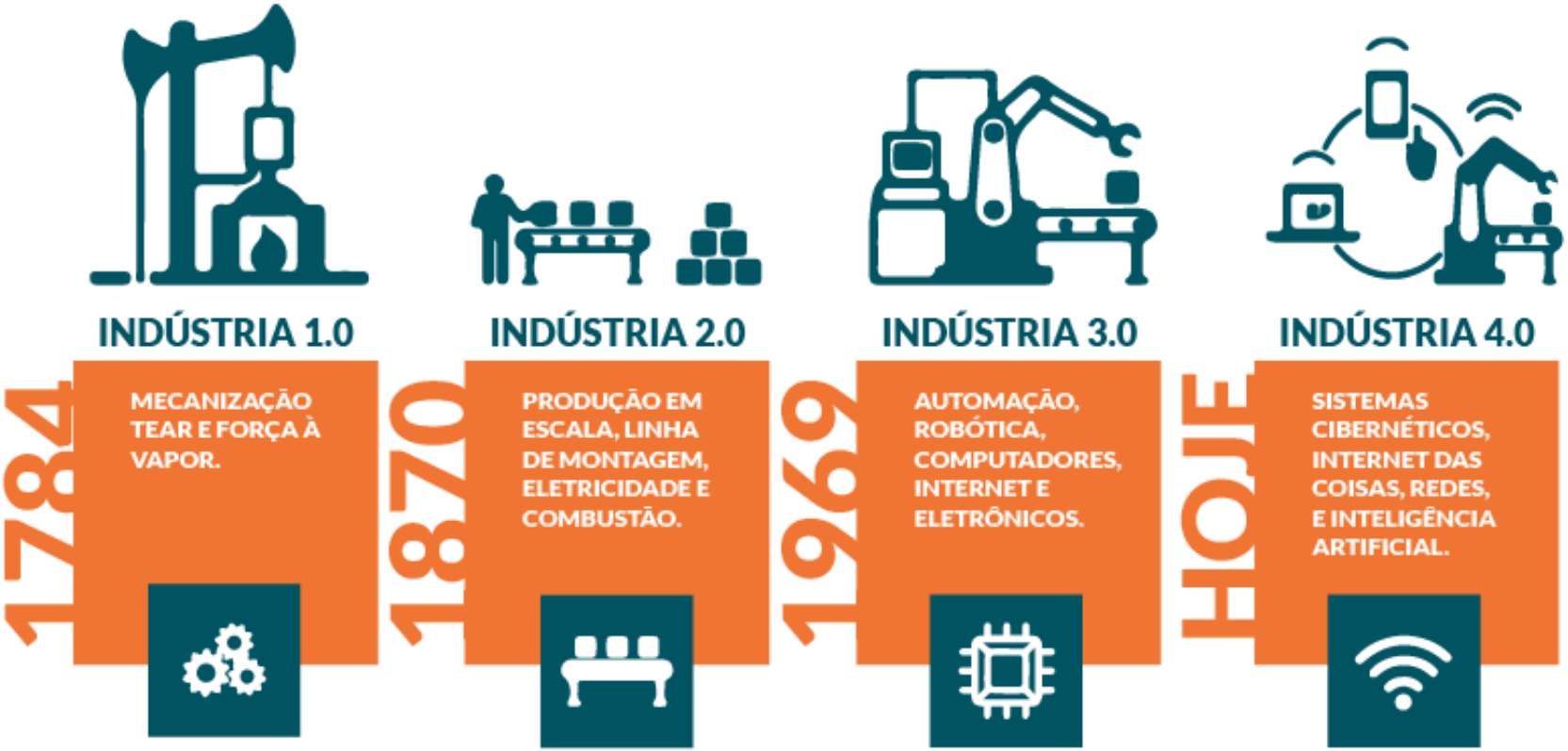
Aula 03

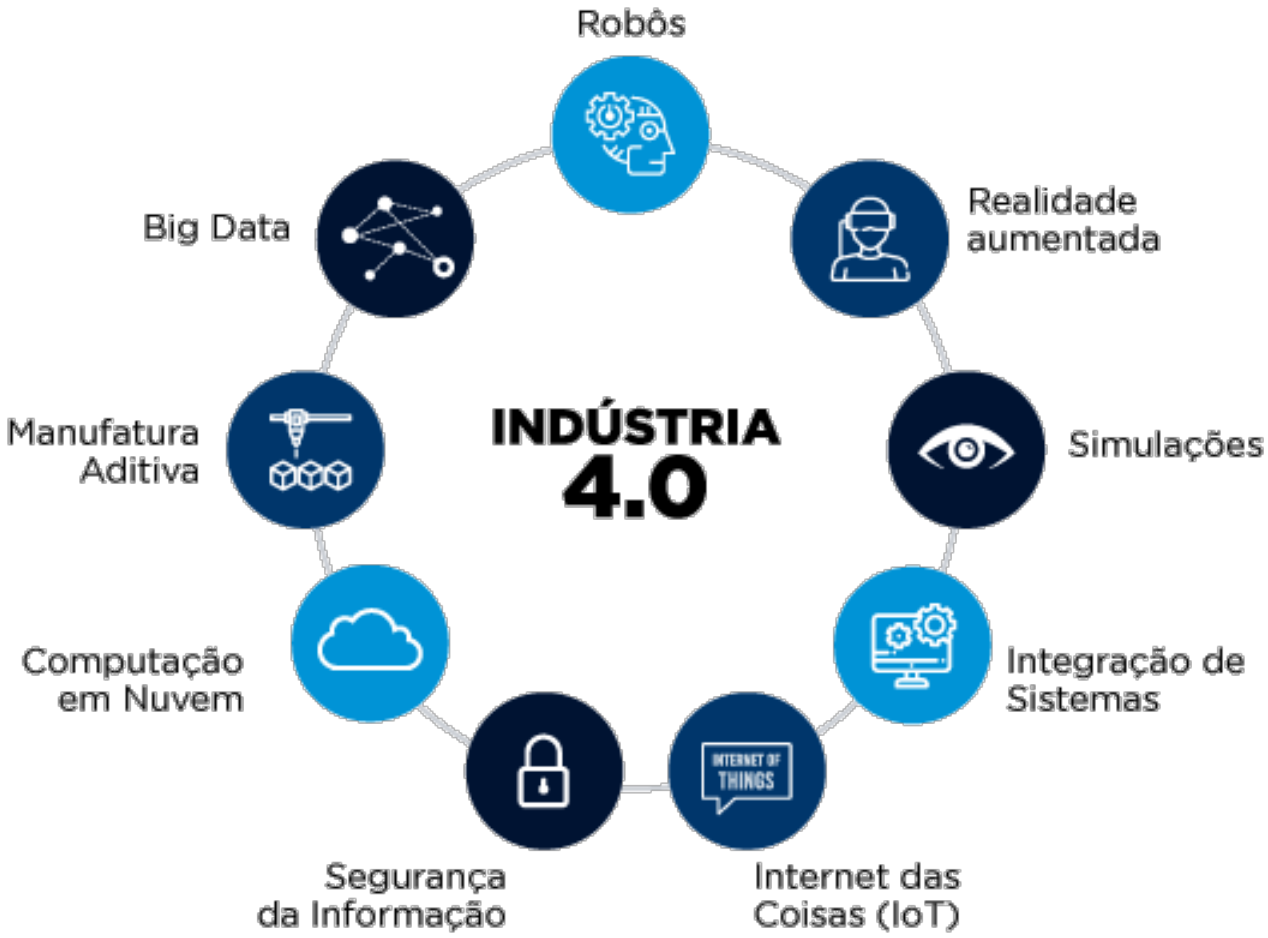
O que é a

INDUSTRIA 4.0

Seu fundamento básico implica que conectando máquinas, sistemas e ativos, as empresas poderão criar redes inteligentes ao longo de toda a cadeia de valor que podem controlar os módulos da produção de forma autônoma.

O que é a Indústria 4.0?



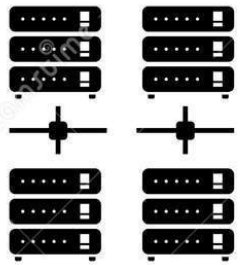

Robôs
Realidade aumentada
Simulações
Integração de Sistemas
Internet das Coisas (IoT)
Segurança da Informação
Computação em Nuvem
Manufatura Aditiva
Big Data
INDÚSTRIA 4.0
INTERNET OF THINGS

BIG DATA

>>

Os sistemas informáticos que existem hoje em dia, os computadores de elevada capacidade e as redes de comunicação abrangentes e de baixo custo, fazem com que seja possível armazenar com rapidez uma grande quantidade de informação, que depois de tratada e analisada em tempo real, facilitará tomar decisões com base nessa informação de valor com mais precisão e confiança.

VOLUME EM BIG DATA



- Como **armazenar** grandes volumes de dados?
- Como **acessar** grandes volumes de dados?
- Precisamos armazenar todos os dados que geramos?

VOLUME EM BIG DATA

Os dados podem ser **ESTRUTURADOS**?

SIM

Data Warehouse

NÃO

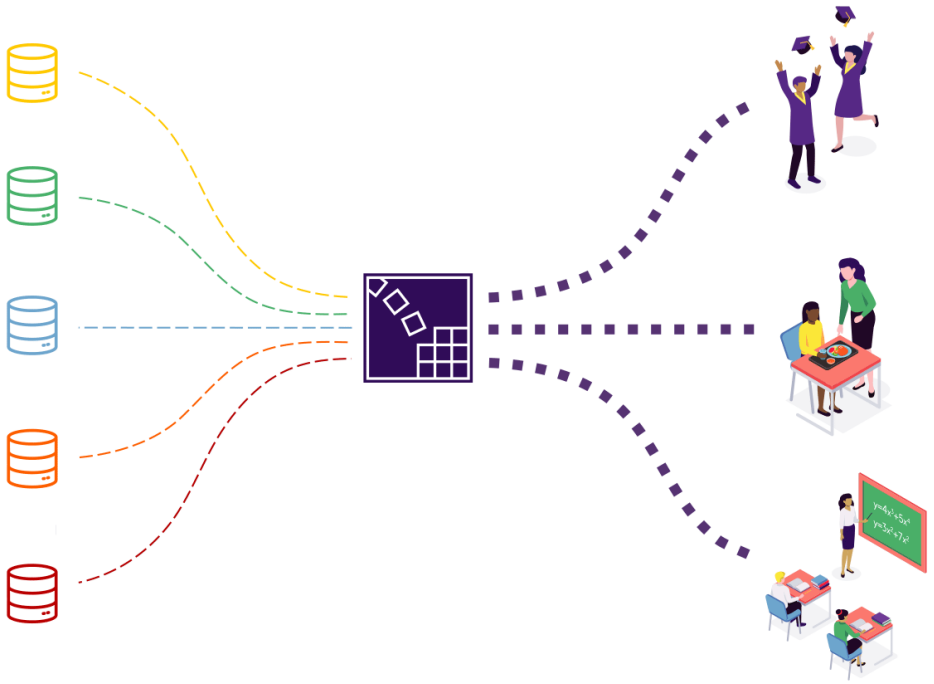
Data Lake ou Data Store

VOLUME EM BIG DATA

Com o Data Warehouse, os dados são limpo e organizados em um único esquema, antes do armazenamento.

A análise é feita consultando diretamente no DW

VOLUME EM BIG DATA



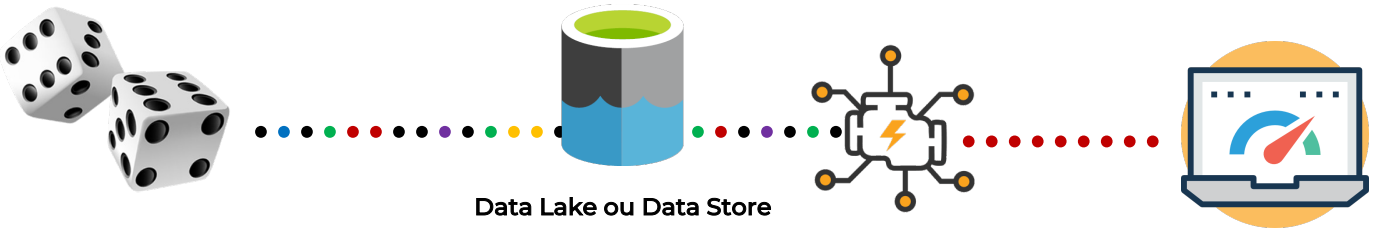
VOLUME EM BIG DATA

Data Lake ou Data Store

Com o Data Lake, os dados são armazenados em seu formato bruto.

Os dados são selecionados e organizados de acordo com a necessidade.

VOLUME EM BIG DATA



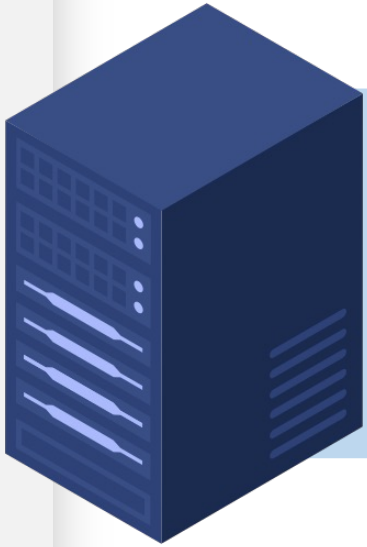
VOLUME EM BIG DATA

Escalabilidade e flexibilidade

ARMAZENAMENTO E PROCESSO PARALELO

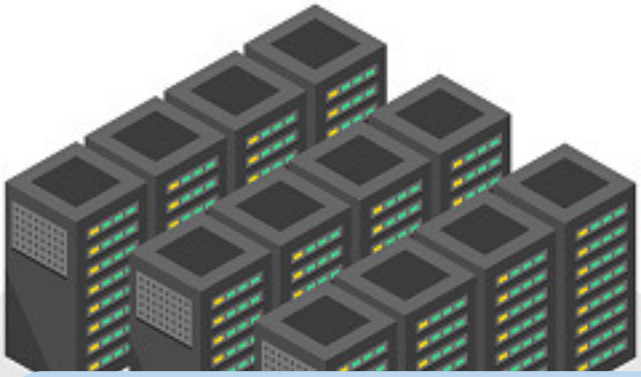
É possível trabalhar com armazenamento e processamento de um BIGDATA em uma única máquina?

SERVIDORES



Um servidor possui escalabilidade vertical, ou seja, há um limite até onde podemos incluir mais espaço em disco, armazenamento, processadores, ...

CLUSTER DE COMPUTADORES



Conjunto de servidores com um mesmo propósito visando fornecer um determinado serviço.

Um CLUSTER possui escalabilidade horizontal, ou seja, se quisermos aumentar a capacidade computacional basta adicionar novos servidores

ARMAZENAMENTO PARALELO

Consiste em distribuir o armazenamento de dados entre diversos servidores permitindo um aumento considerável utilizando hardware de baixo custo.

BIG DATA

Como gerenciamos o Armazenamento Paralelo?

Por meio do **sistema de arquivos distribuídos**

Prof. Leandro de Souza

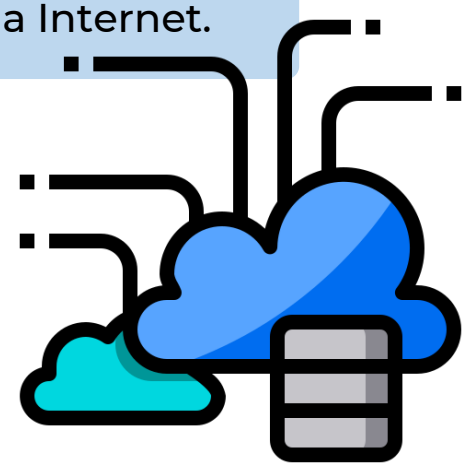
BIG DATA

CLOUD COMPUTING

Prof. Leandro de Souza

O QUE É CLOUD COMPUTING?

É a tecnologia que permite o uso remoto de recursos da computação por meio da conectividade da Internet.



Fornecedores

- AWS (Amazon Web Services)
- Azure (Microsoft)
- Google Cloud
- IBM Cloud


DESAFIADORES
LÍDERES
Amazon Web Services
Microsoft
Google
Alibaba Cloud
Oracle
Tencent Cloud
IBM
OPERADORES DE NICHO
VISIONÁRIOS
CAPACIDADE DE EXECUÇÃO
ABRANGÊNCIA DA VISÃO
Em julho 2021
© Gartner, Inc

Como vamos Processar os Dados?

Por meio do **PROCESSAMENTO PARALELO**

É um método que permite que dois (ou mais) processadores trabalhem em partes diferentes de uma determinada tarefa para obter resultados mais rápidos.

O objetivo é dividir uma tarefa em várias sub-tarefas e executá-las em paralelo.

Processamento Paralelo com

Ao utilizarmos um framework de processamento paralelo, as sub-tarefas são levadas para o processador da máquina do cluster em que os dados estão armazenados, aumentando assim a velocidade de processamento de grandes volumes de dados

BIG DATA

APACHE HADOOP HDFS

Hadoop Distributed File System

Ele processa eficientemente grandes volumes de dados em um cluster de hardware commodity

Prof. Leandro de Souza

APACHE HADOOP HDFS

- Blocos de dados
- Tamanho fixo: 128 MB
- Um arquivo muito grande pode ter blocos armazenados em mais de um servidor

APACHE HADOOP HDFS

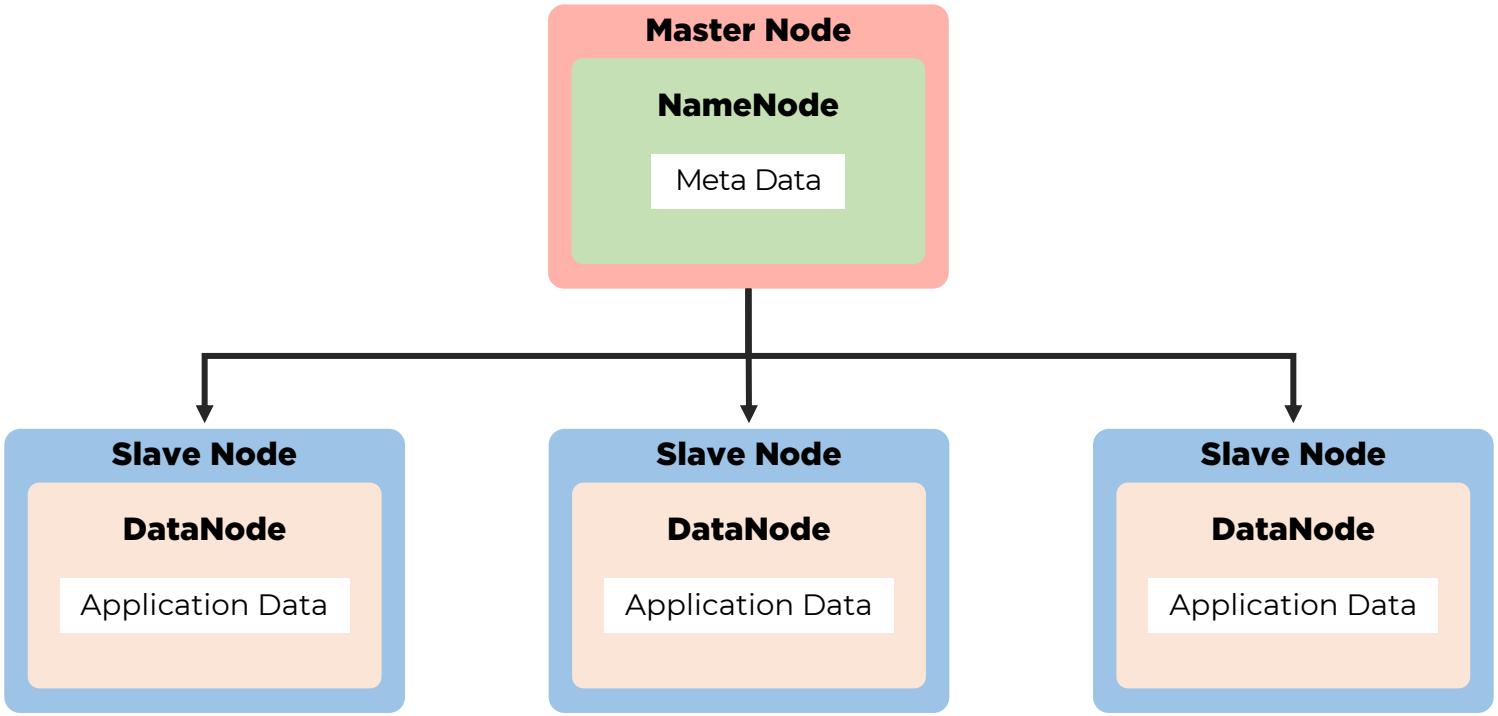
Arquivo de 64 MB

Desperdício de 64 MB

APACHE HADOOP HDFS

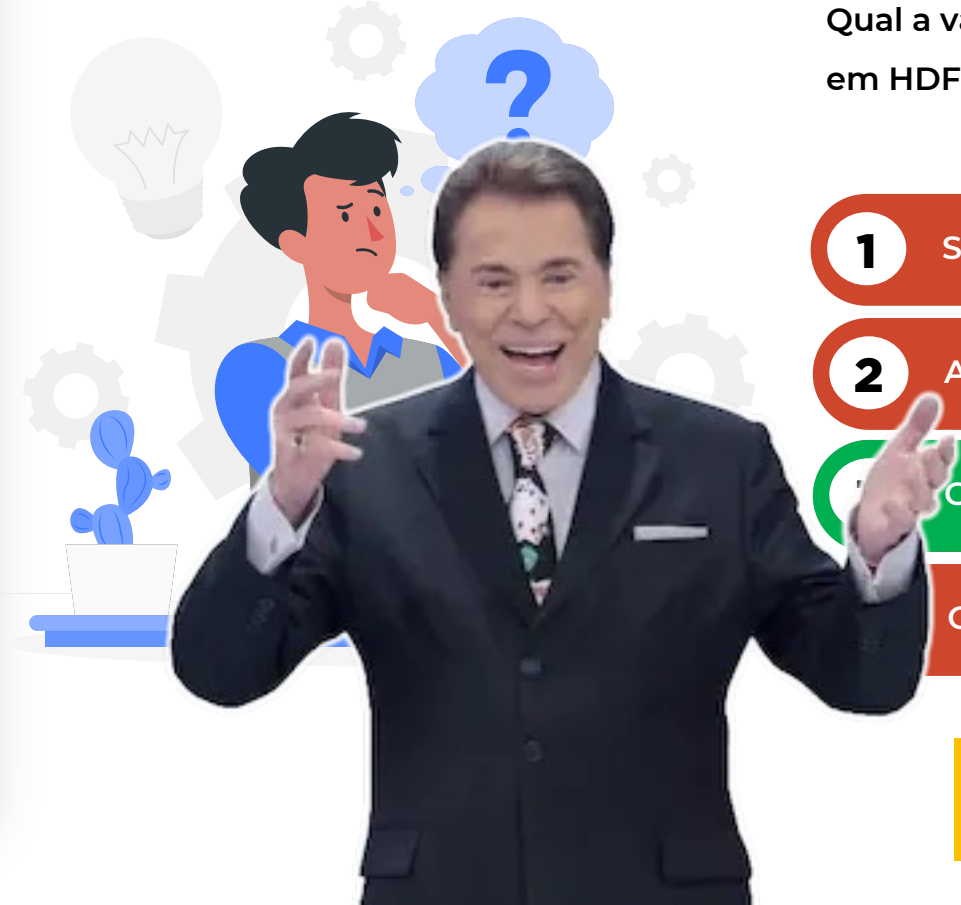
- Sistema de arquivos distribuído
- Composto por NameNodes e DataNodes
- Alto desempenho
- Integridade
- Prioriza consistência e tolerância a falhas
- Arquivos divididos em blocos
- Tamanho padrão do bloco (64 ou 128MB)
- Dados replicados
- NameNode é único ponto de falha

Arquitetura base do hadoop



SISTEMAS OPERACIONAIS

Para pensar...



Qual a vantagem do sistema de replicação 3x em HDFS?

1. Suporta processamento paralelo
2. Análise de dados mais rápida
3. Garantir a tolerância a falhas
4. Gerencia recursos em cluster

Prof. Leandro de Souza