

Aprendizado de Máquina com Dados em Larga Escala

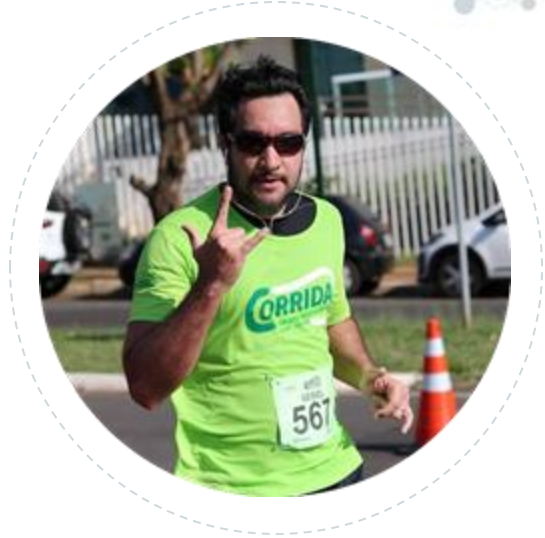
Gesiel Rios Lopes¹

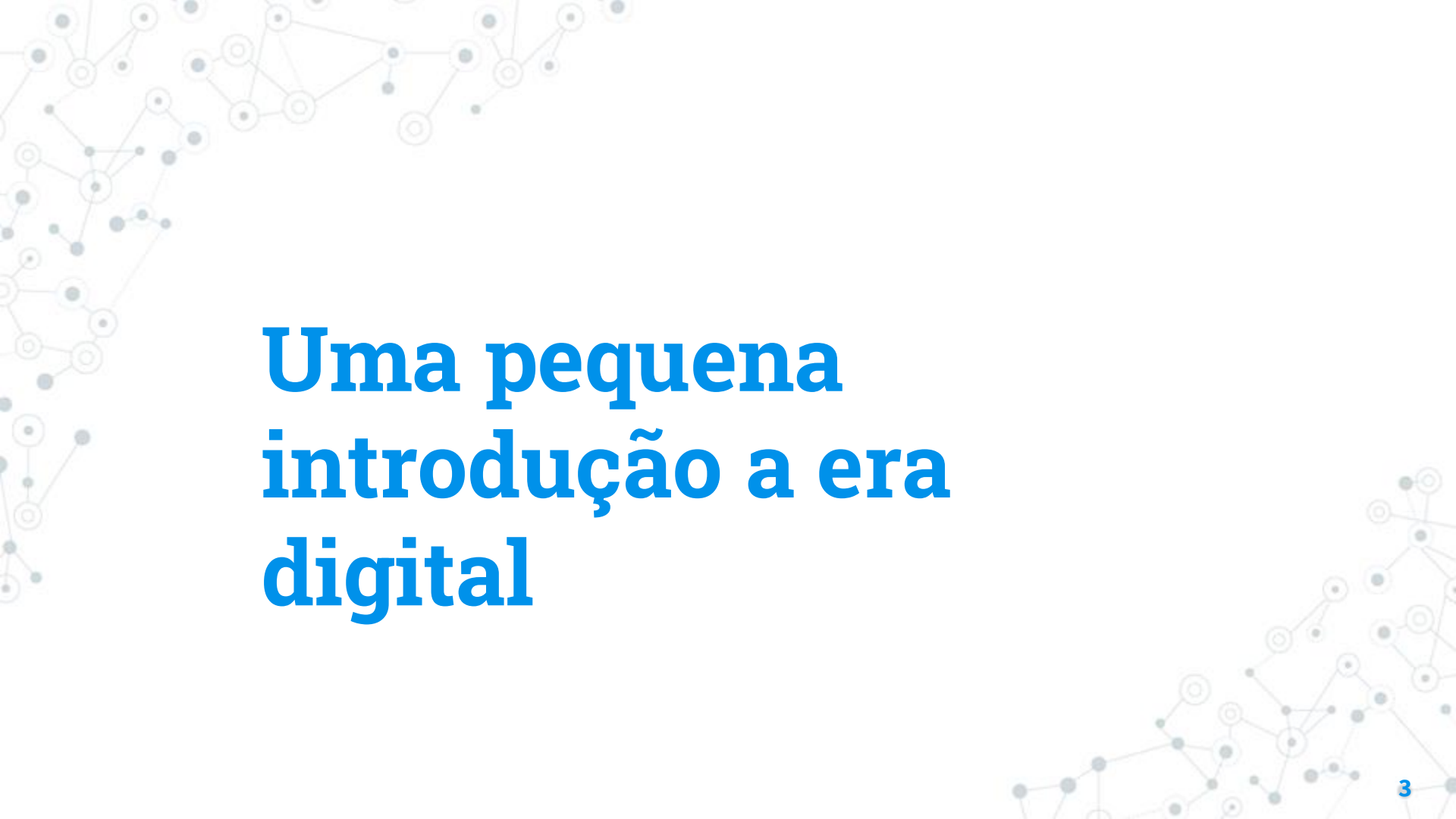
¹Laboratório de Sistemas Embarcados e Evolutivos (LSEE)
Instituto de Ciências Matemática e de Computação (ICMC)
Universidade de São Paulo (USP)

gesielrios@usp.br

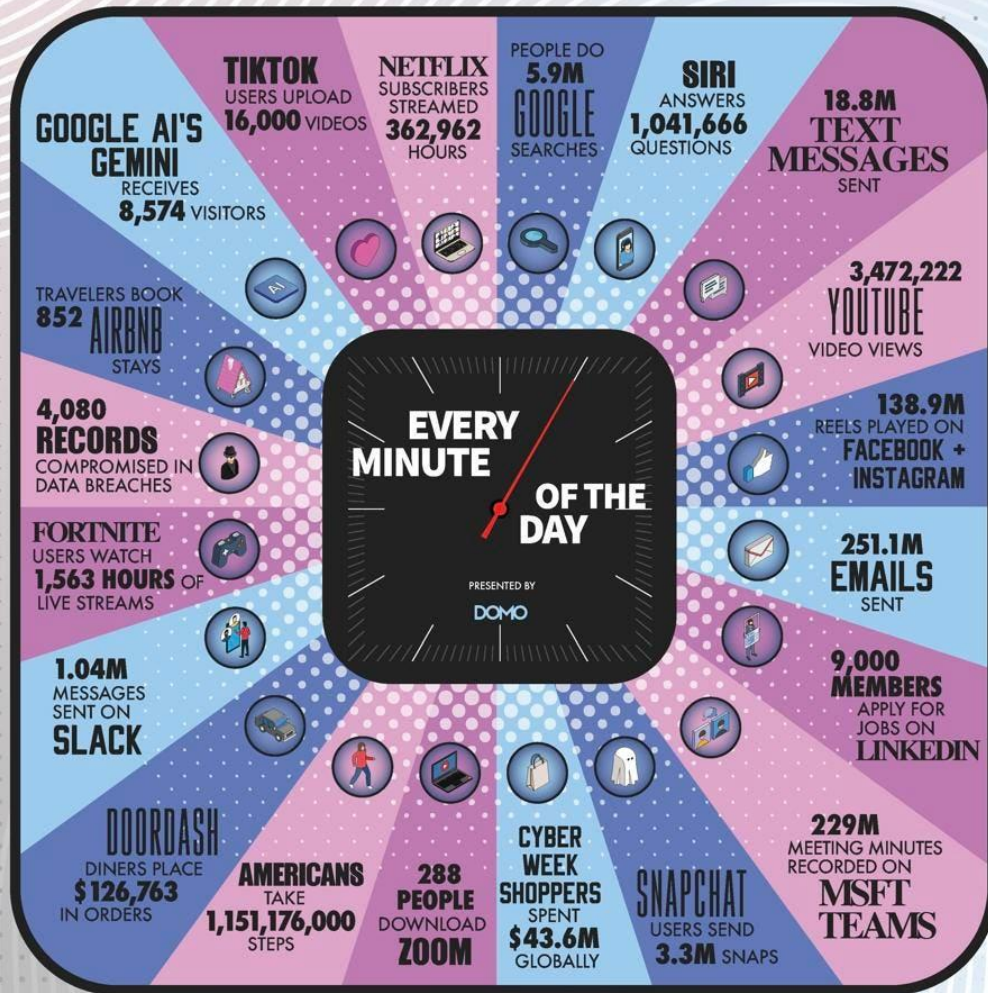
About me, by myself

Gesiel Rios Lopes



A decorative background featuring a network diagram with nodes and connecting lines, primarily located in the top-left and bottom-right corners.

Uma pequena introdução a era digital

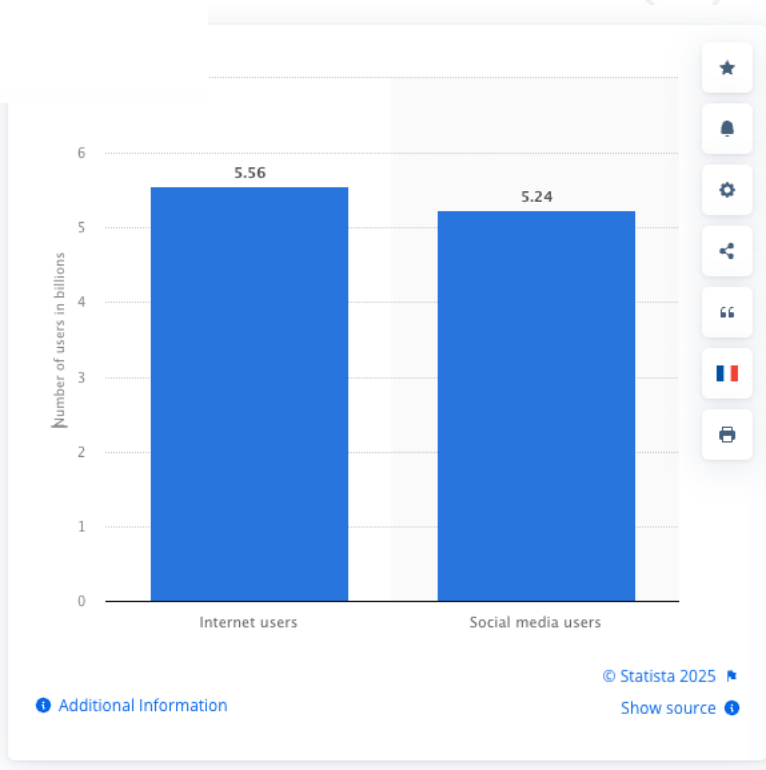


Fonte:

<https://www.forbes.com/sites/timbajarin/2025/01/21/mind-blowing-stats-of-what-transpires-on-the-internet-every-minute/>

Number of internet and social media users worldwide as of February 2025

(in billions)



Fonte: <https://www.statista.com/statistics/617136/digital-population-worldwide/>

Introdução

- © Vivemos em uma época digital onde existe uma enorme quantidade de dados sendo produzidas, onde a quantidade de informações produzidas atualmente é superior à quantidade de informação que uma pessoa conseguiria lidar durante toda sua vida (Marquesone, 2016).
- © Este cenário proporciona um mundo cheio desafios importantes no armazenamento, manipulação e análise de forma inteligente (Goldman et al., 2012).

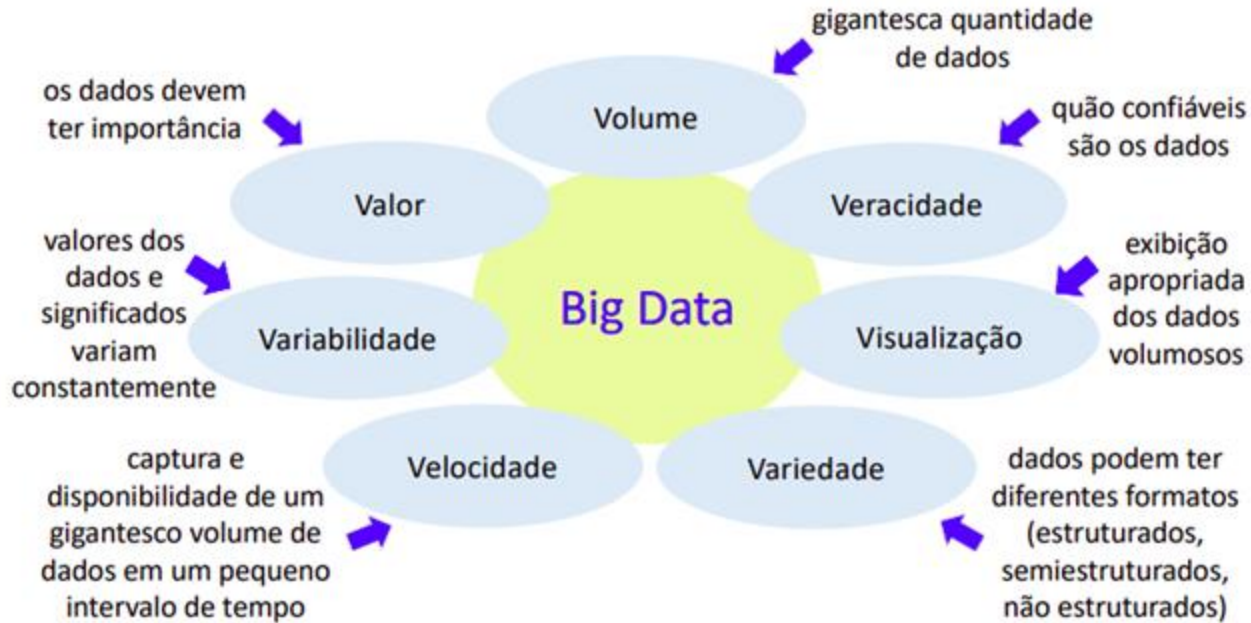
Introdução

- ◎ Baseado nessas e em outras aplicações que possuem um volume gigantesco de dados, surgiu o conceito denominado “Big Data”.
- ◎ Esse termo faz menção não apenas ao volume, mas também à sua variedade e a velocidade necessária para o seu processamento.

Introdução

- ◎ Big Data é um termo usado principalmente para descrever os conjuntos de dados que são muito grandes e complexos e que requerem de tecnologias avançadas de armazenamento, gestão, análise e visualização (Marquesone, 2016).

Introdução



Introdução




48%

dos entrevistados acreditam
que dados e informações são
relevantes na tomada de decisão

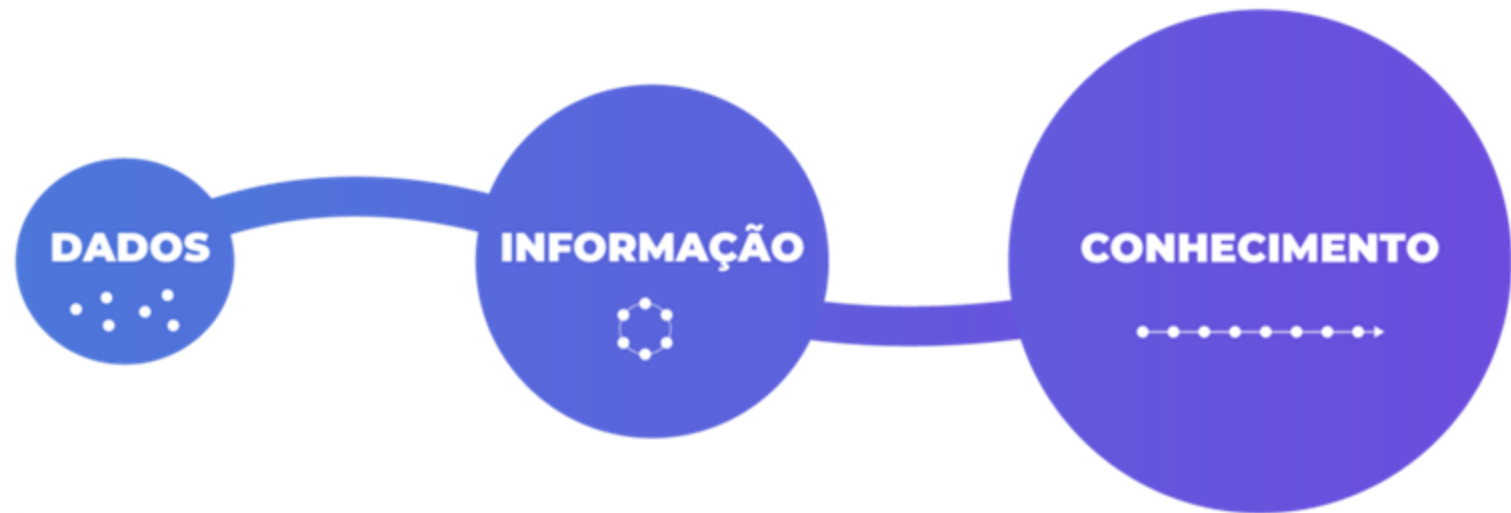
67%

dos entrevistados acreditam
que, no futuro, dados e informações
serão relevantes na tomada de
decisão

FONTE: [HTTPS://BI-SURVEY.COM/](https://bi-survey.com/)



Introdução



Introdução

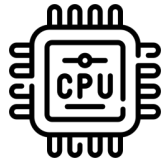
- ◎ Big Data é um conceito abrangente, que engloba ferramentas, técnicas e até mesmo define características dos dados
- ◎ O mundo está se preparando cada vez mais para o uso de soluções de Big Data, devido ao seu enorme potencial de geração de valor para o negócio
- ◎ Apesar desse potencial, o uso de Big Data traz contrapartidas que devem ser levadas em consideração

A decorative background featuring a network diagram with nodes and connecting lines, primarily located in the top-left and bottom-right corners.

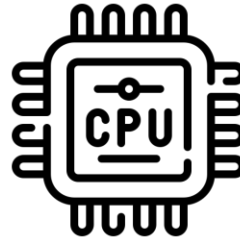
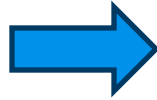
Processamento de Dados Massivos

O limite físico dos hardwares

- Até meados dos anos 2000, os computadores desenvolvidos até então avançavam em termos de hardware no seguinte sentido:



Processador de
1 só núcleo

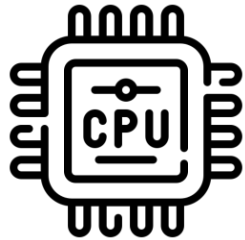


Processador mais
poderoso de 1 só núcleo

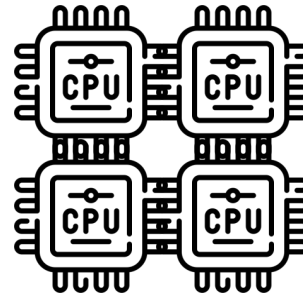
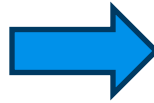
- É possível entender isso como uma escalabilidade vertical, porque o que é feito é aumentar significativamente a capacidade de um único componente.

O limite físico dos hardwares

- © No entanto, esse modelo de escalabilidade vertical atingiu um limite físico: chegou um ponto em que não era possível dissipar a quantidade de calor gerada pelos processadores.



Processador mais poderoso de 1 só núcleo



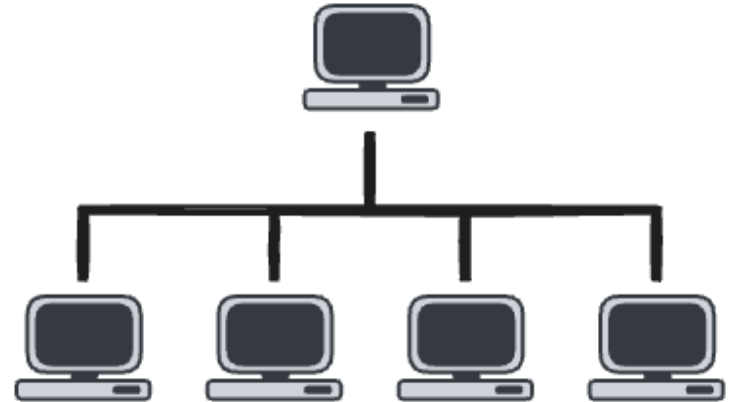
Processador com vários Núcleo menos poderosos

Processamento Paralelo

- © Com a disponibilidade de várias unidades de processamento, é possível adotar a estratégia de **paralelização** das tarefas da seguinte forma:
 1. Programa recebe uma tarefa de grande custo computacional.
 2. A tarefa grande é quebrada em várias listas de pequenas tarefas, que são enviadas aos executores.
 3. Cada executor realiza suas tarefas ao mesmo tempo, uma de cada vez.
 4. Os resultados das tarefas dos executores são agregados para retornar o resultado final.

Processamento Distribuído

- © O **processamento distribuído** segue os mesmos princípios da computação em paralelo com a diferença de que a paralelização é feita a partir de uma rede de computadores interligados (também chamado de cluster), ao invés de ser somente a nível de processadores de uma máquina.



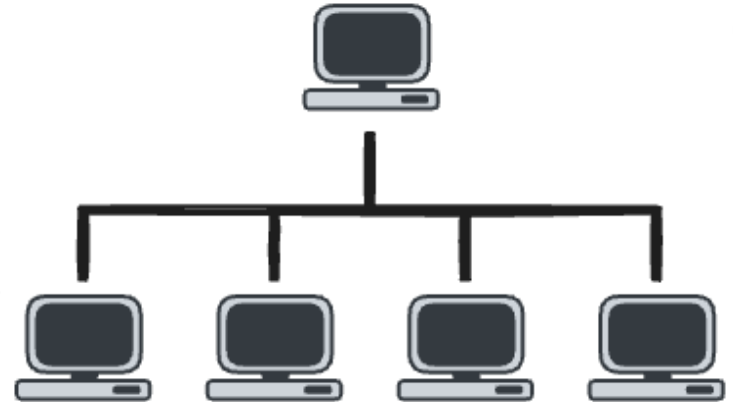
Processamento Distribuído

⊙ Vantagens da Paralelização

- Escalabilidade
- Tolerância a Falhas

⊙ Desafios

- Gerenciamento de Tarefas
- Gerenciamento de Partições de Dados
- Segurança
- Gerenciamento de Executores

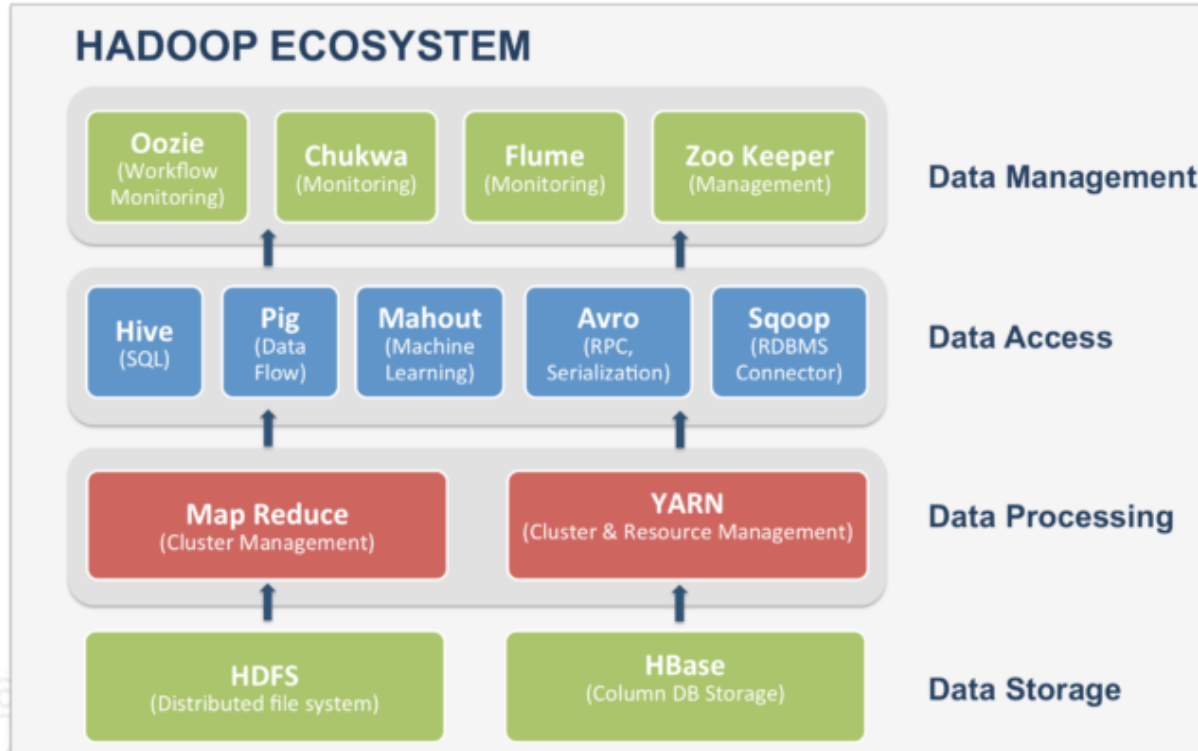


Ecossistema Apache Hadoop

- ◎ Arcabouço para processamento e armazenamento de dados em larga escala:
 - Código aberto
 - Implementado em Java
 - Inspirado no GFS e MapReduce do Google
 - Projeto top-level da Fundação Apache
 - Tecnologia recente, porém já muito utilizada



Ecossistema Apache Hadoop



Ecossistema Apache Hadoop

- ◎ A biblioteca principal conta com os seguintes módulos:
 - Hadoop Common: módulo que contém os utilitários comuns a todos os outros módulos do Hadoop.
 - Hadoop Distributed File System (HDFS): módulo que contém as funcionalidades relacionadas ao armazenamento distribuído de dados.
 - Hadoop MapReduce: módulo que oferece serviços de computação distribuída no ambiente Hadoop.
 - Hadoop YARN: módulo que realiza o gerenciamento de recursos e divisão de tarefas dentro do ambiente distribuído do Hadoop.



Ecossistema Apache Hadoop

- © Além desses módulos, o framework conta com tecnologias disponíveis para outros propósitos, como:
 - Apache HIVE: banco de dados que utiliza uma interface de SQL no ambiente distribuído.
 - Apache Mahout: módulo para a criação de aplicações de machine learning.



Ecosystem Apache Hadoop

- © Além desses módulos, o framework conta com tecnologias disponíveis para outros propósitos, como:
 - Apache Ambari: serviços de provisionamento, gerenciamento e monitoramento de clusters no Apache Hadoop.
 - Apache Oozie: serviços de agendamento de jobs.
 - Apache Zookeeper: módulo para coordenar os serviços do Ecosystem Hadoop.



Ecossistema Apache Hadoop

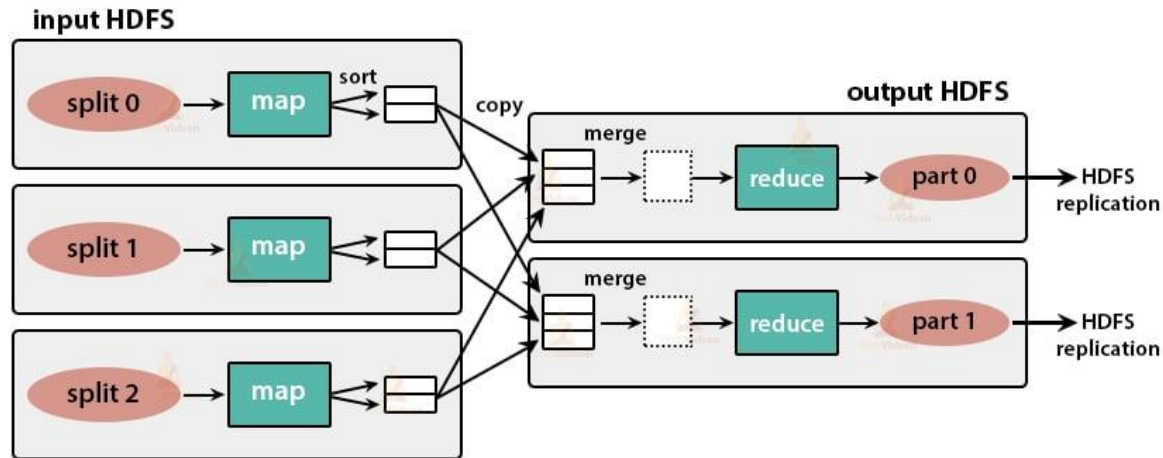
◎ MapReduce

- O Hadoop MapReduce assim como a implementação anterior do Google, é a ferramenta utilizada no processamento paralelo e distribuído de volumes massivos de dados
- Ela funciona com base em duas operações distintas:
 - ◎ **Map:** usada para segmentar os dados em pares chave/valor antes do processamento
 - ◎ **Reduce:** Operação responsável por realizar as computações em forma de operações de agregação



Ecossistema Apache Hadoop

Apache Hadoop MapReduce



Ecossistema Apache Hadoop

© Desvantagens do MapReduce

- Complexidade Operacional
- Modelo de Programação Verboso
- Escrita em Disco





Introdução ao Apache Spark

O que é o Apache Spark?

- ◎ O Apache Spark é um framework 100% open source de processamento distribuído e computação em clusters, projetado especialmente para trabalhar com quantidades massivas de dados.
- ◎ O framework dispõe de vários componentes para diferentes formas de processamento dos dados, como operações em dados estruturados, processamento em streaming, computação de grafos e até mesmo ajuste de modelos de Machine Learning.



Principais Características

1

Velocidade

Simplicidade

2

3

Modularidade

Extensibilidade

4



Principais Características

◎ Velocidade

- Armazenamento em memória.
- Directed Acyclic Graph (DAG).
- Catalyst Optimizer.
- Lazy Evaluation.

	Hadoop MR Record	Spark Record	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
# Nodes	2100	206	190
# Cores	50400 physical	6592 virtualized	6080 virtualized
Cluster disk throughput	3150 GB/s (est.)	618 GB/s	570 GB/s
Sort Benchmark Daytona Rules	Yes	Yes	No
Network	dedicated data center, 10Gbps	virtualized (EC2) 10Gbps network	virtualized (EC2) 10Gbps network
Sort rate	1.42 TB/min	4.27 TB/min	4.27 TB/min
Sort rate/node	0.67 GB/min	20.7 GB/min	22.5 GB/min

Fonte: <https://www.databricks.com/blog/2014/11/05/spark-officially-sets-a-new-record-in-large-scale-sorting.html>



Principais Características

◎ Simplicidade

- No seu nível mais baixo, o Spark abstrai os dados em uma estrutura chamada de **Resilient Distributed Dataset (RDD)**, que é uma coleção particionada e imutável de registros.
- Operar com RDDs é uma tarefa bastante complexa, e, por isso, nas APIs mais recentes do Spark, é possível realizar operações com essas construções a partir dos Spark DataFrames, uma estrutura baseada em linhas e colunas que é bastante comum no contexto de dados.



Principais Características

© Simplicidade

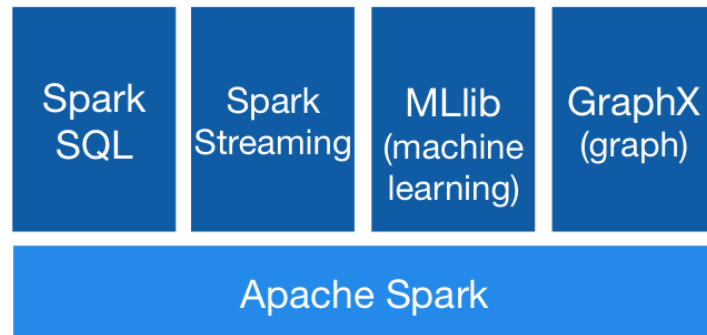
- Além disso, as operações sobre dados estruturados herdam princípios fortes do SQL, e é, inclusive, possível utilizar a popular linguagem de consultas para manipular os dados.



Principais Características

© Modularidade

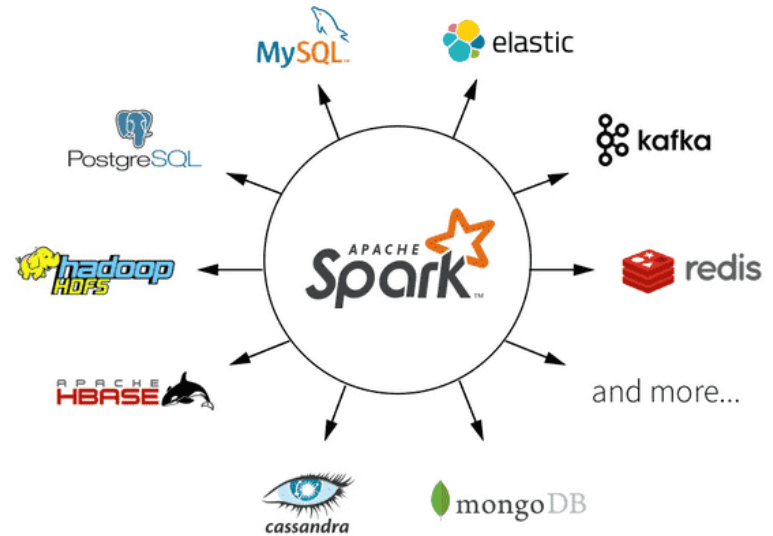
- Spark SQL: Módulo para o processamento de dados estruturados.
- Structured Streaming: Módulo para o processamento de dados em tempo real.
- Spark ML: Módulo de pré-processamento e Machine Learning.
- GraphX: Módulo para computação de grafos.



Principais Características

◎ Extensibilidade

- Spark é unicamente uma ferramenta de processamento de dados distribuídos e, por causa disso, é capaz de se conectar com as mais diferentes fontes de dados.

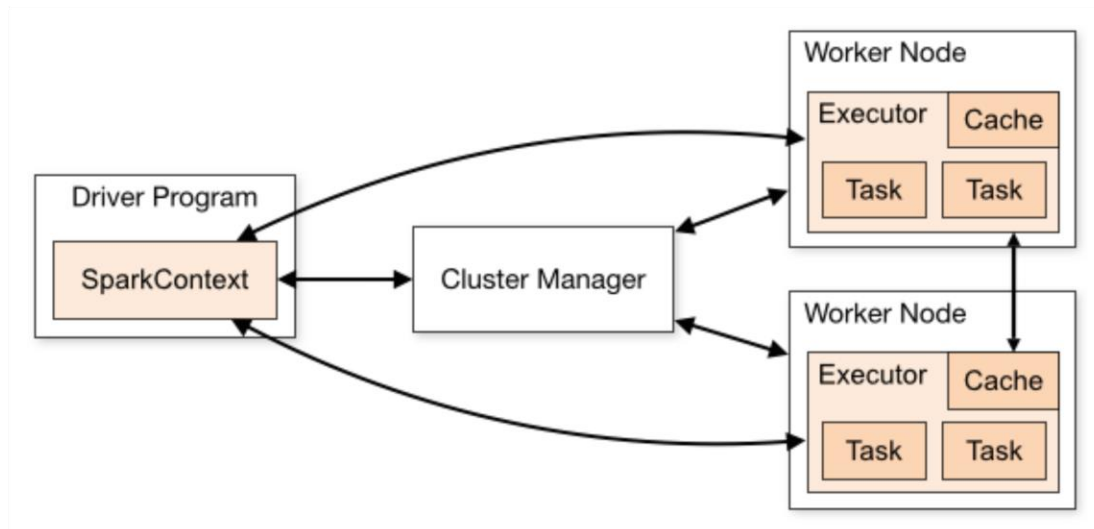


Aplicação Spark

- © A arquitetura distribuída do Spark é dividida em dois componentes principais:
 - Driver: Responsável por manter informações sobre a aplicação, responder a inputs do programa ou usuário e analisar, distribuir e agendar tarefas nos executores.
 - Executores: São responsáveis por, de fato, realizar as tarefas designadas a eles pelo driver.



Aplicação Spark



Spark Session

- ◎ A Spark Session é o ponto de entrada pra acessar todas as funcionalidades do Spark. Por meio dela, é possível:
 - Ler e criar DataFrames;
 - Realizar queries do SQL;
 - Configurar a aplicação;
 - Acessar o catálogo de metadados.
- ◎ Além disso, esse objeto também unifica o ponto de entrada de todos os módulos, não só do Spark SQL.





Obrigado!

Perguntas ?



github.com/gesielrios



gesielrios@gmail.com



<https://www.linkedin.com/in/gesiel-rios-lopes-815801244>



That's all Folks!