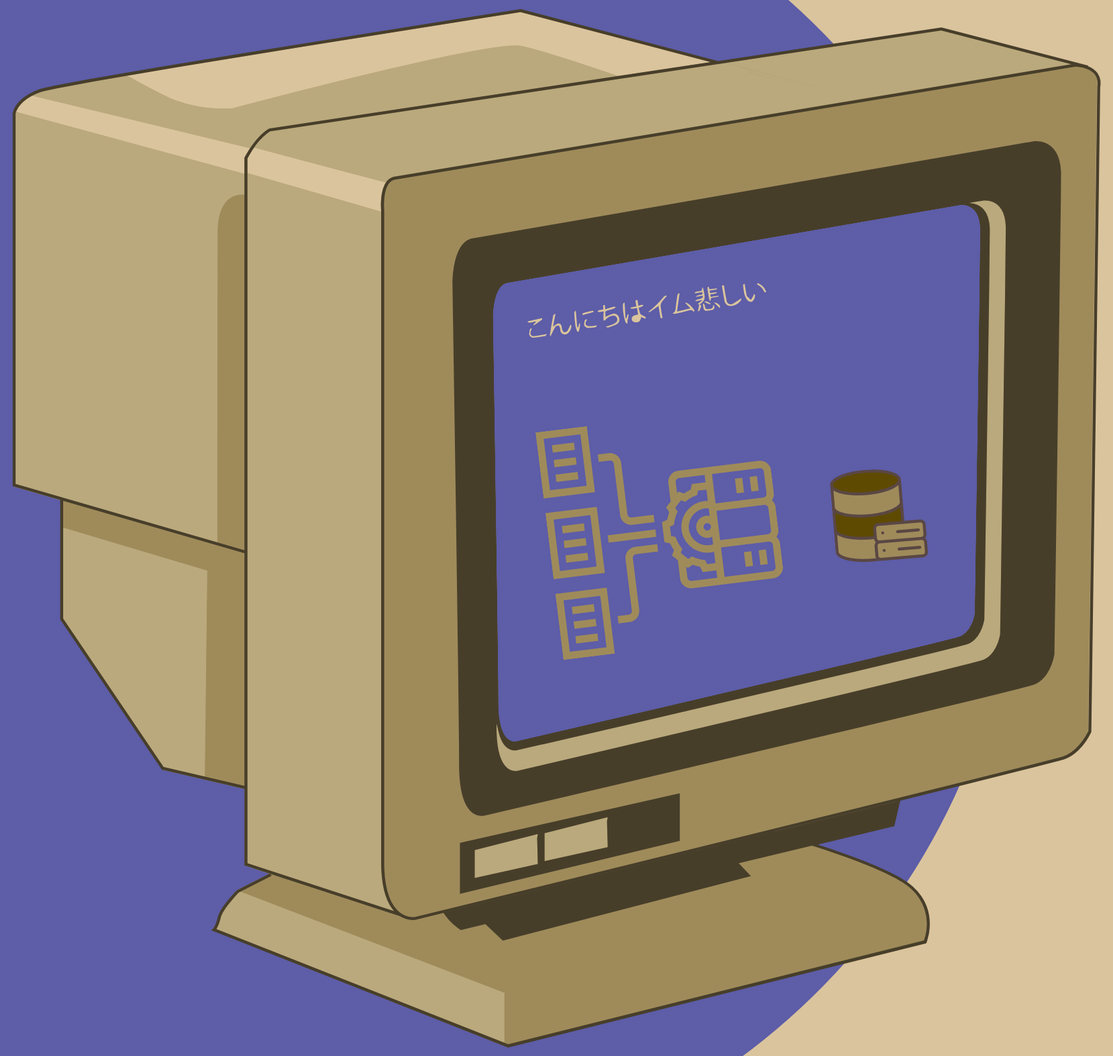


ARQUITETURA DE ELT E ETL PARA INTEGRAÇÕES CORPORATIVAS

MBA EM BUSINESS INTELLIGENCE E BIG DATA

Prof. Thiago Santos

UNIPÊ - 2025

AULA - 01

SOBRE MIM ...

8 Anos ná área de Dados

5 Anos trabalhando com projetos exclusivamente de Engenharia de Dados

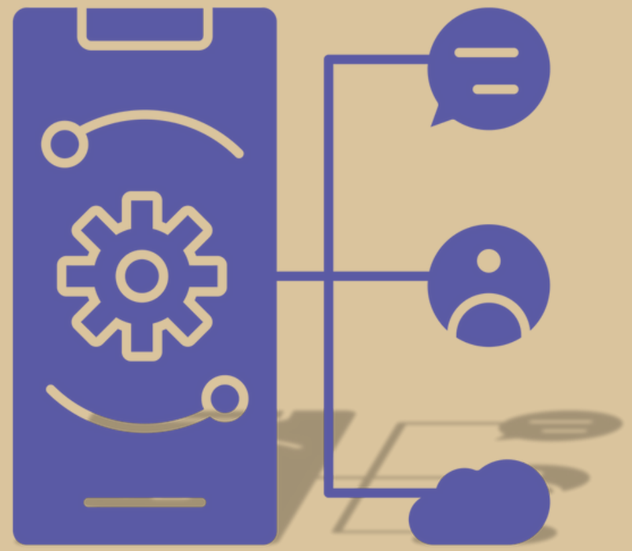
Certificados (Airflow - Databricks - AWS - Kafka)

WWW.THIS-THIAGO.MEDIUM.COM

O QUE ESPERAR?

Você aprenderá os conceitos fundamentais de arquiteturas de ELT e ETL, com foco em integrações corporativas. Será apresentado o manejo de dados estruturados, semi-estruturados e não estruturados, além de práticas com ferramentas como Apache Spark, Hadoop, e MinIO.

Você também explorará os conceitos de virtualização e conteinerização, tecnologias fundamentais para o gerenciamento e escalabilidade de aplicações em ambientes corporativos.



AGENDA

TEORIA

1. Dados estruturados, semi-estruturados e não estruturados
2. ETL & ELT
3. Traditional Data Warehouse (DW)
4. Fundamentos de Virtualização e Conteinerização
5. Processamento de dados com Apache Spark

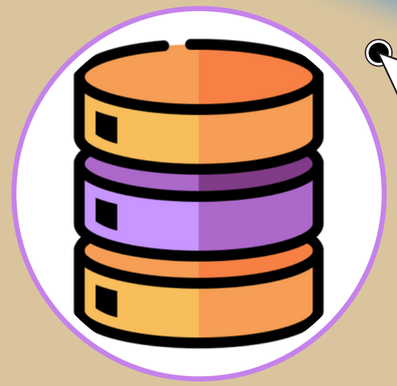
AGENDA

PRÁTICA

1. Ingestão de dados com Apache Spark
2. Tratamento de dados com Dataframe API
3. Traditional Data Warehouse (DW)
4. Tratamento de dados com SparkSQL
5. Jupyter + Pyspark + MinIO + Postgres + S3 + Document DB

AULA 01

DADOS ESTRUTURADOS



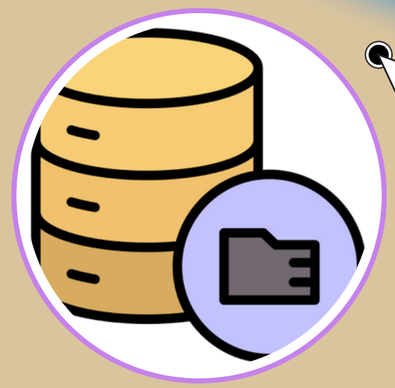
Dados estruturados são dados com um formato padronizado para acesso eficiente por software e humanos e com um modelo definido na sua concepção.

- Modelo de dados
- Inflexíveis
- Consolidados

AWS.AMAZON.COM

AULA 01

DADOS NÃO ESTRUTURADOS



Os dados não estruturados são informações sem nenhum modelo de dados definido, geralmente porque é ineficiente modelá-los.

- **Imprevisíveis**
- **Análise mais complexa**
- **Ausência de modelo de dados**

AWS.AMAZON.COM

AULA 01

DADOS SEMI-ESTRUTURADOS

São organizados, mas não seguem uma estrutura de esquema definida, ao invés disso são organizados por meio de etiquetas ou "tags".

- Auto Descritivos
- Flexíveis
- Ausência de modelo de dados

AWS.AMAZON.COM

AULA 01

ETL - EXTRACT TRANSFORM AND LOAD

Correspondem ao processo de extração, transformação (limpeza e combinação) de dados de várias fontes e carregamento em um outro repositório para serem usados em análises.

Com a aplicação do processo de extração, transformação e carregamento (ETL), conjuntos de dados brutos individuais podem ser preparados em um formato e uma estrutura mais consumíveis para fins de análise, resultando em informações mais significativas.

AWS.AMAZON.COM

AULA 01

ELT - EXTRACT LOAD AND TRANSFORM

Inverte a ordem das operações do ETL onde os dados são carregados diretamente no sistema de destino antes de qualquer processamento.

Com a evolução das tecnologias de nuvem, as empresas passaram a armazenar dados brutos ilimitados em grande escala e analisá-los conforme o necessário.

AWS.AMAZON.COM

AULA 01

TRADITIONAL DATA WAREHOUSE (DW)

BILL INMON

Define um Data Warehouse como um repositório centralizado de dados, construído para apoiar a tomada de decisão, segue uma abordagem top-down.

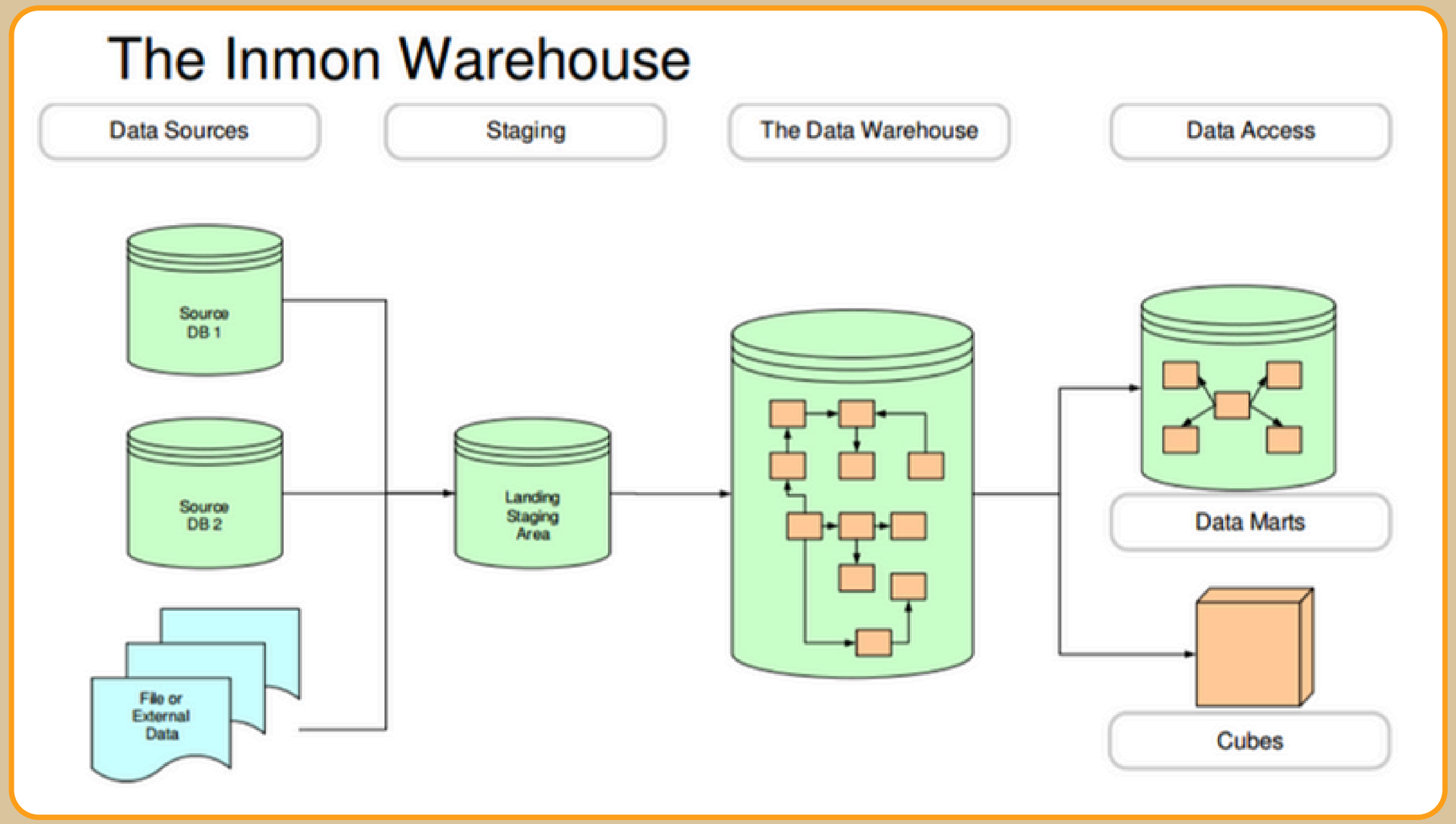
RALPH KIMBALL

DW é construído de forma incremental, começando com data marts integrados que fornecem uma visão lógica e consolidada dos dados para suporte à decisão, uma abordagem bottom-up.

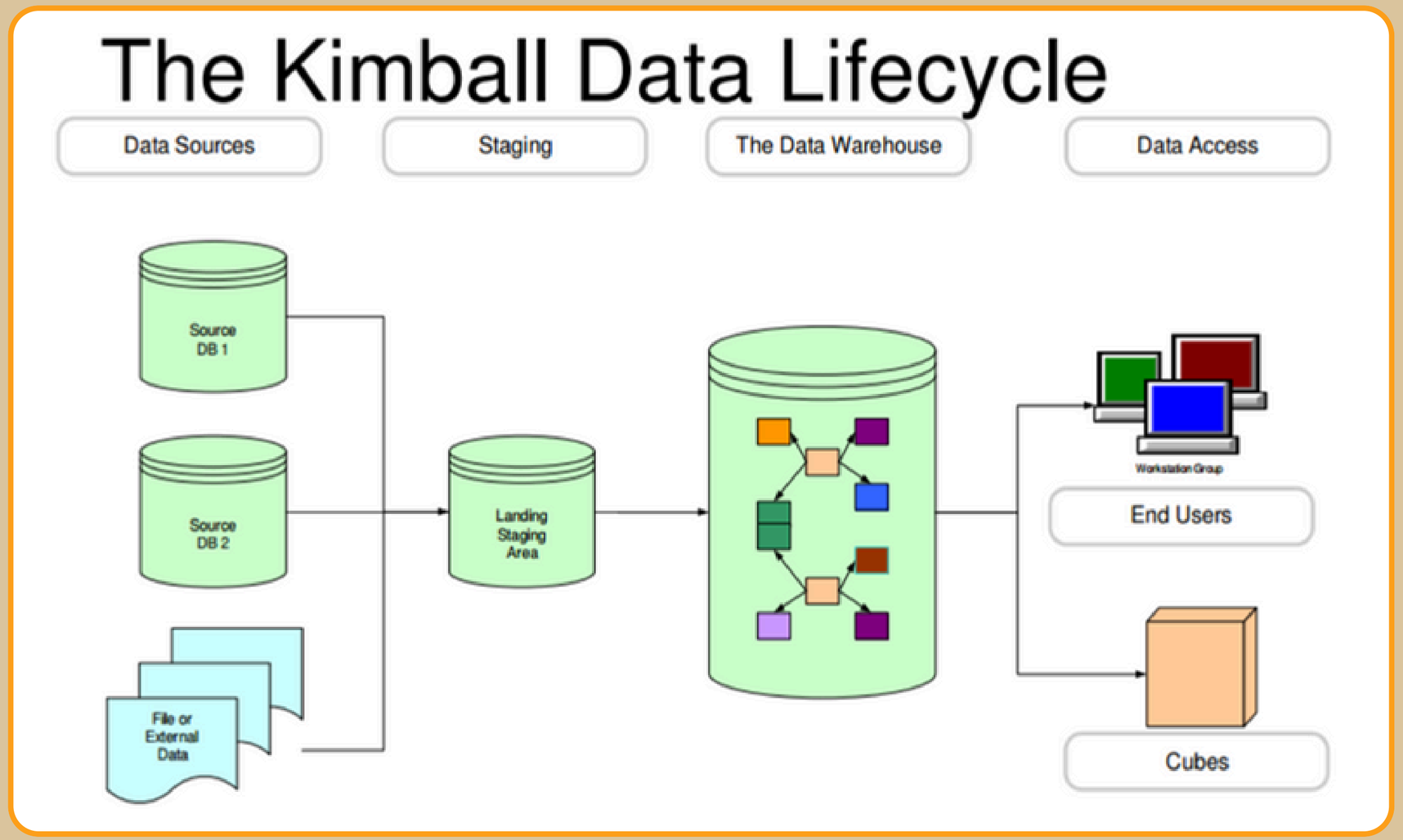
AWS.AMAZON.COM

AULA 01

ABORDAGEM TOP-DOWN - CENTRALIZAÇÃO, GOVERNANÇA E CONSISTÊNCIA

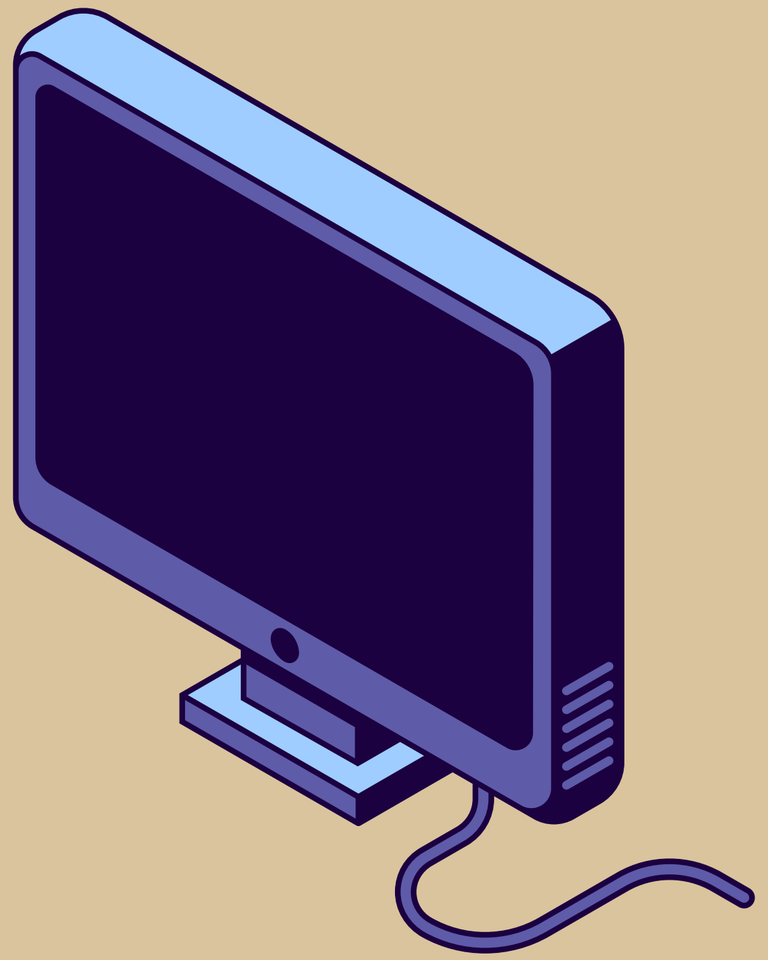


ABORDAGEM BOTTOM-UP - FACILIDADE DE ANÁLISE E RAPIDEZ



AULA 01

VIRTUALIZAÇÃO



A virtualização de servidores é um processo que particiona um servidor físico em vários servidores virtuais. É uma maneira eficiente e econômica de usar recursos de servidor e implantar serviços de TI em uma organização.

- Uso eficiente de recursos
- Gerenciamento automatizado de TI
- Recuperação de desastres mais rápida

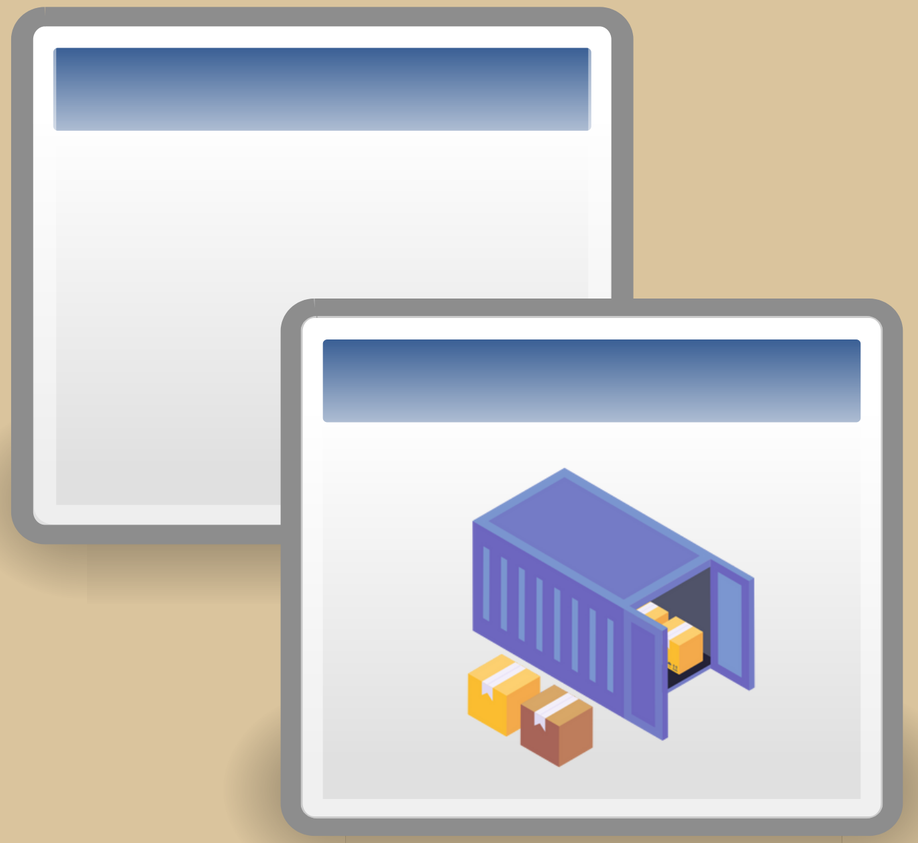
AWS.AMAZON.COM

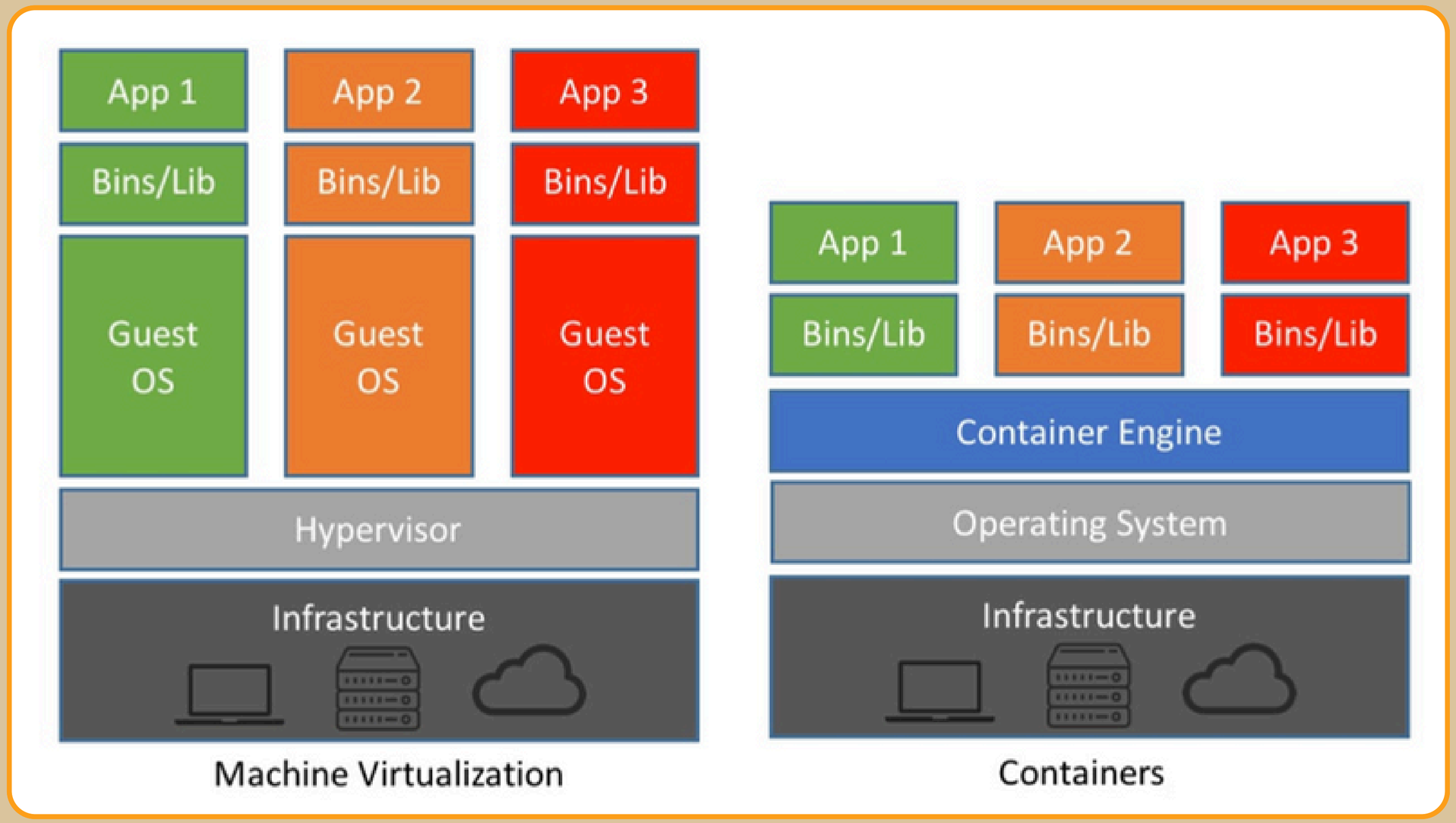
AULA 01

CONTEINERIZAÇÃO

A conteinerização é um processo de implantação de software que agrupa o código de uma aplicação com todos os arquivos e bibliotecas de que ela precisa para ser executado em qualquer infraestrutura.

- **Portabilidade**
- **Escalabilidade**
- **Tolerância a falhas**
- **Agilidade**




App 1
App 2
App 3
Bins/Lib
Bins/Lib
Bins/Lib
Guest OS
Guest OS
Guest OS
Hypervisor
Infrastructure
Machine Virtualization
App 1
App 2
App 3
Bins/Lib
Bins/Lib
Bins/Lib
Container Engine
Operating System
Infrastructure
Containers

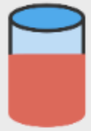
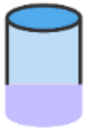
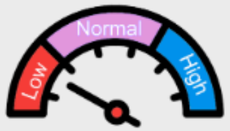
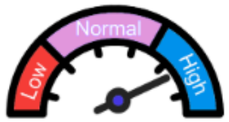
AULA 01

VIRTUALIZAÇÃO

É como se você tivesse vários "computadores virtuais" em um único computador.

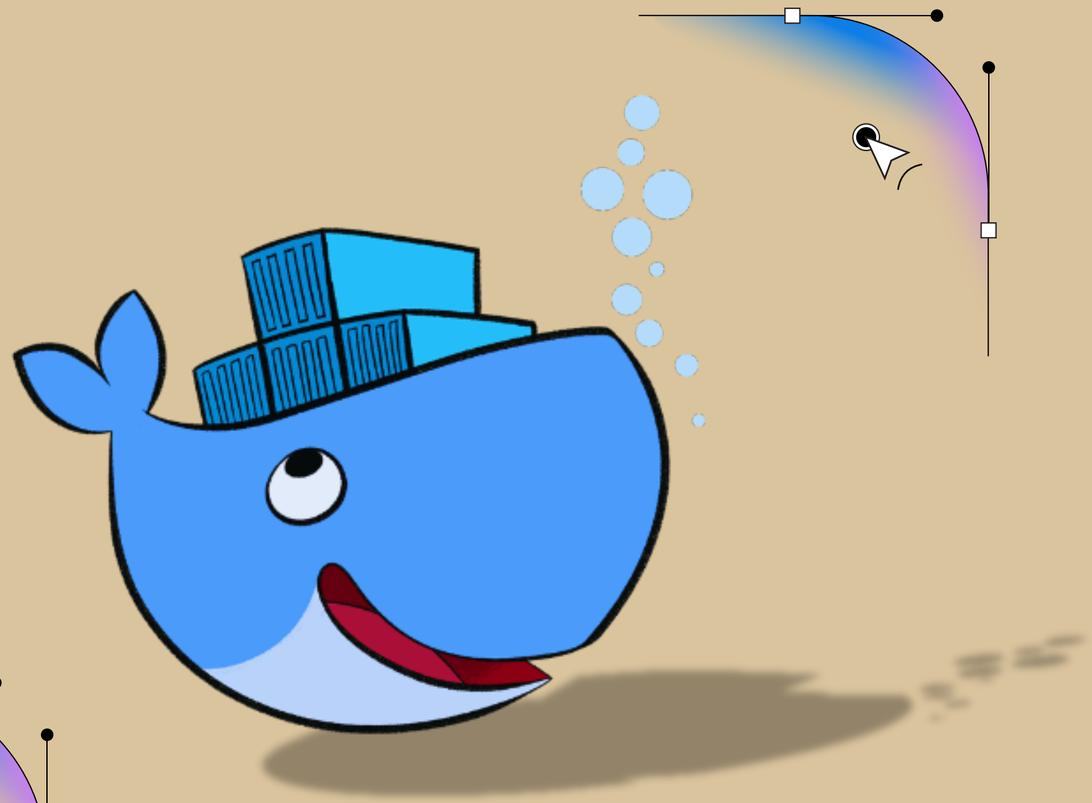
CONTEINERIZAÇÃO

É como se você colocasse uma aplicação e tudo o que ela precisa dentro de uma "caixa", que pode ser transportada e executada em qualquer lugar.

	Virtualization	Containerization
Startup time	 minutes	 seconds
Disk space		
Portability	Less Portable	
Efficiency		
Operating system/kernel	Dedicated	Shared

AULA 01

DOCKER

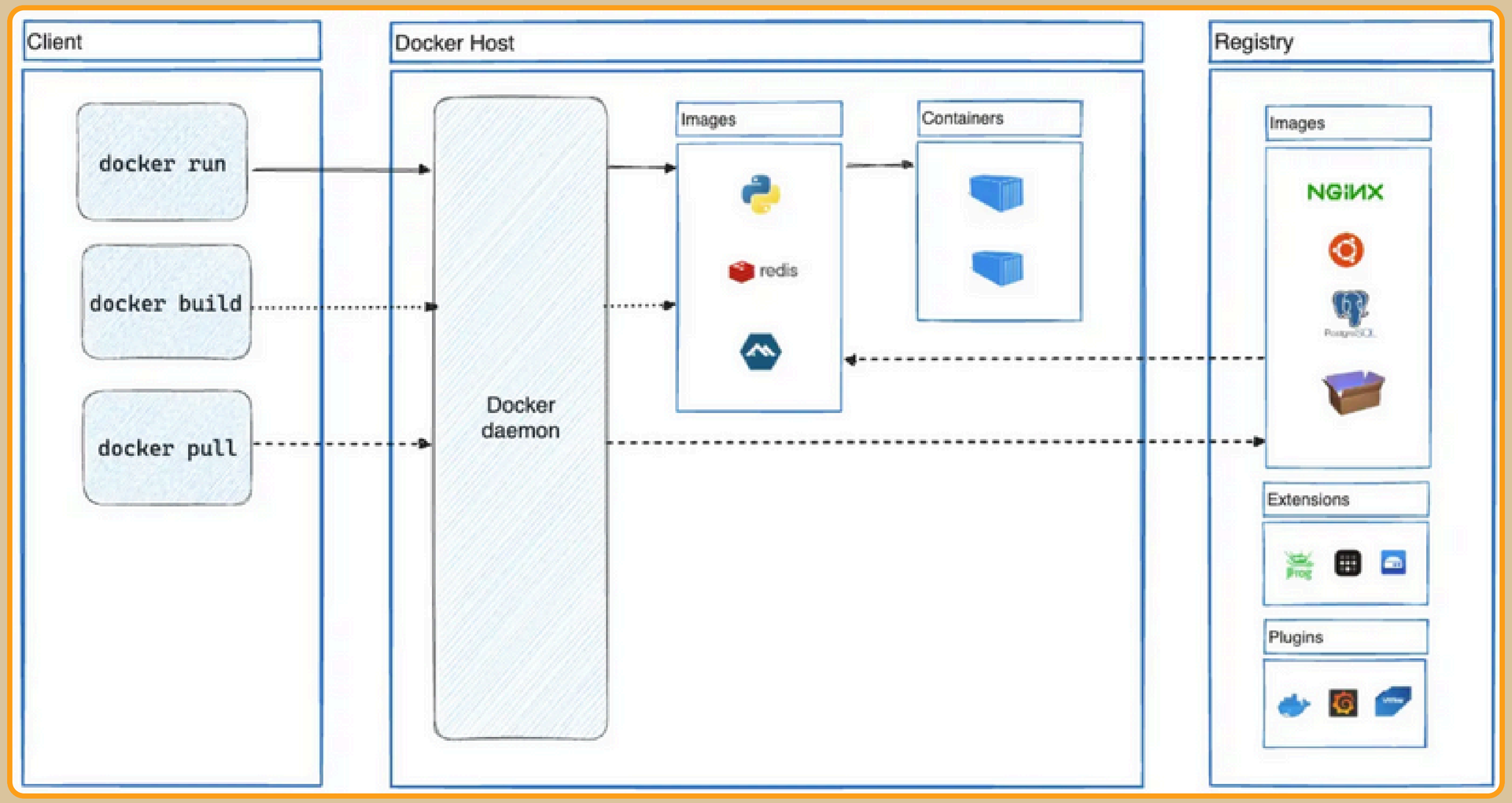


O Docker é uma plataforma aberta para desenvolver, enviar e executar aplicativos. O Docker cria pacotes de software em unidades padronizadas chamadas de conteiners que têm tudo o que o software precisa para ser executado.

- Open Source
- Separa as aplicações da infraestrutura
- Leve e portátil

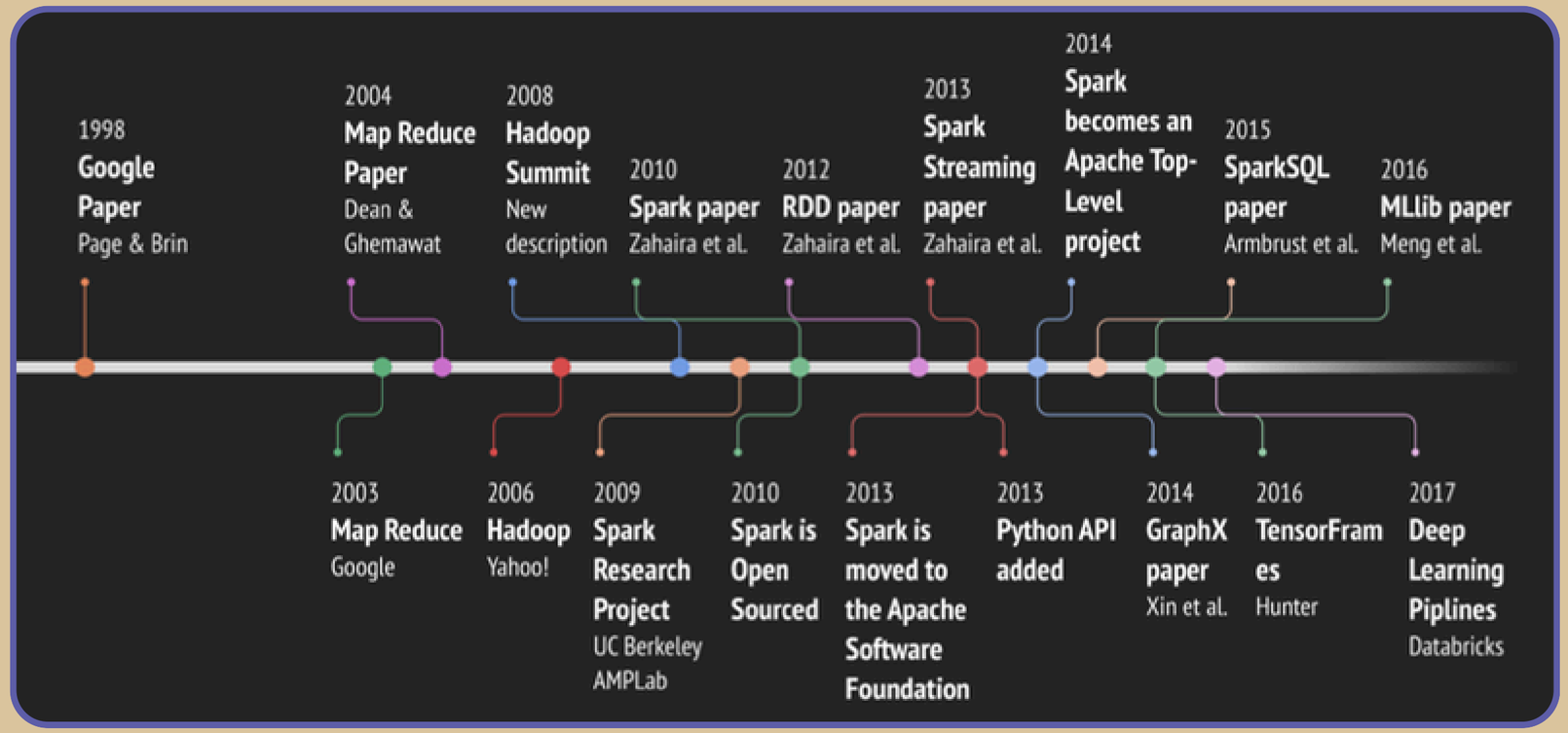
WWW.DOCS.DOCKER.COM

ARQUITETURA DOCKER



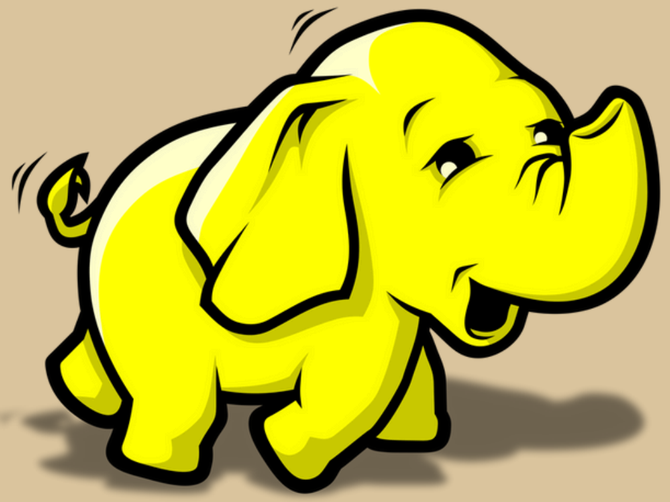
AULA 01

PROCESSAMENTO DE DADOS COM APACHE SPARK



AULA 01

O QUE É O APACHE HADOOP?

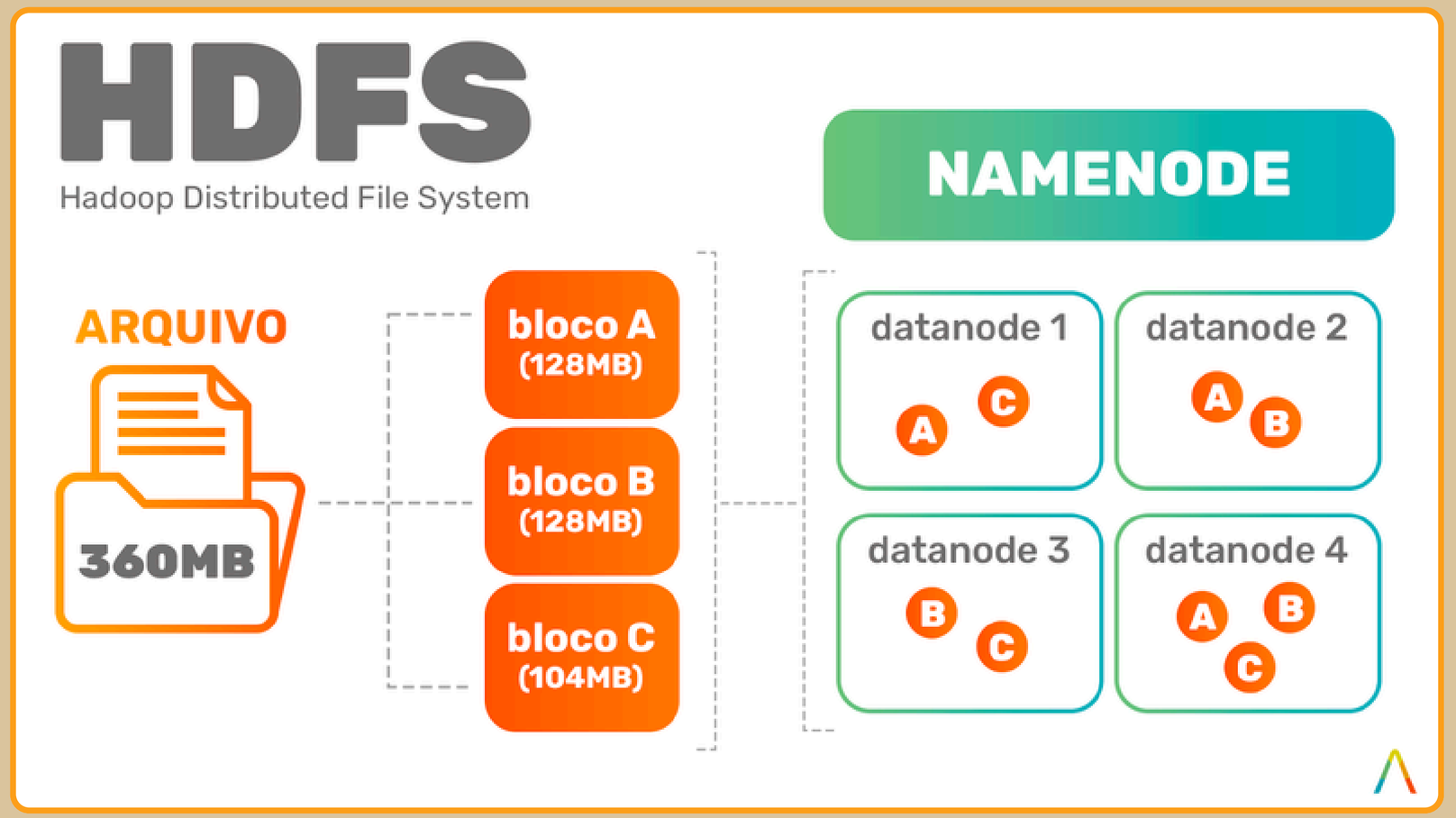


Projeto de software de código aberto que pode ser usado para processar de modo eficiente grandes conjuntos de dados.

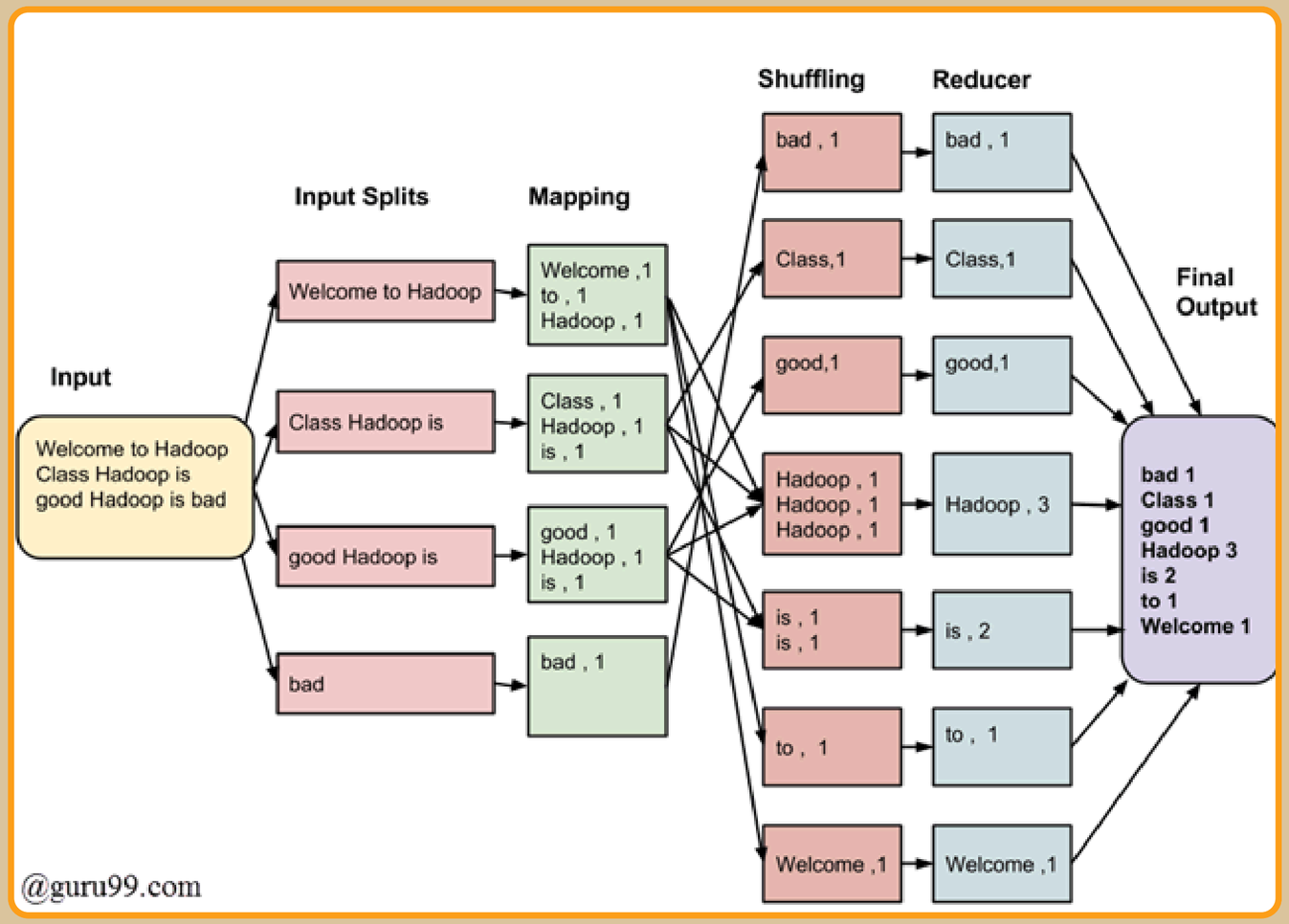
- **Armazenamento distribuído**
- **MapReduce**
- **Escalável**

AWS.AMAZON.COM

HDFS


HDFS
Hadoop Distributed File System
NAMENODE
ARQUIVO
360MB
bloco A
(128MB)
bloco B
(128MB)
bloco C
(104MB)
datanode 1
A
C
datanode 2
A
B
datanode 3
B
C
datanode 4
A
B
C

EXEMPLO PROCESSO DE MAP REDUCE



AULA 01

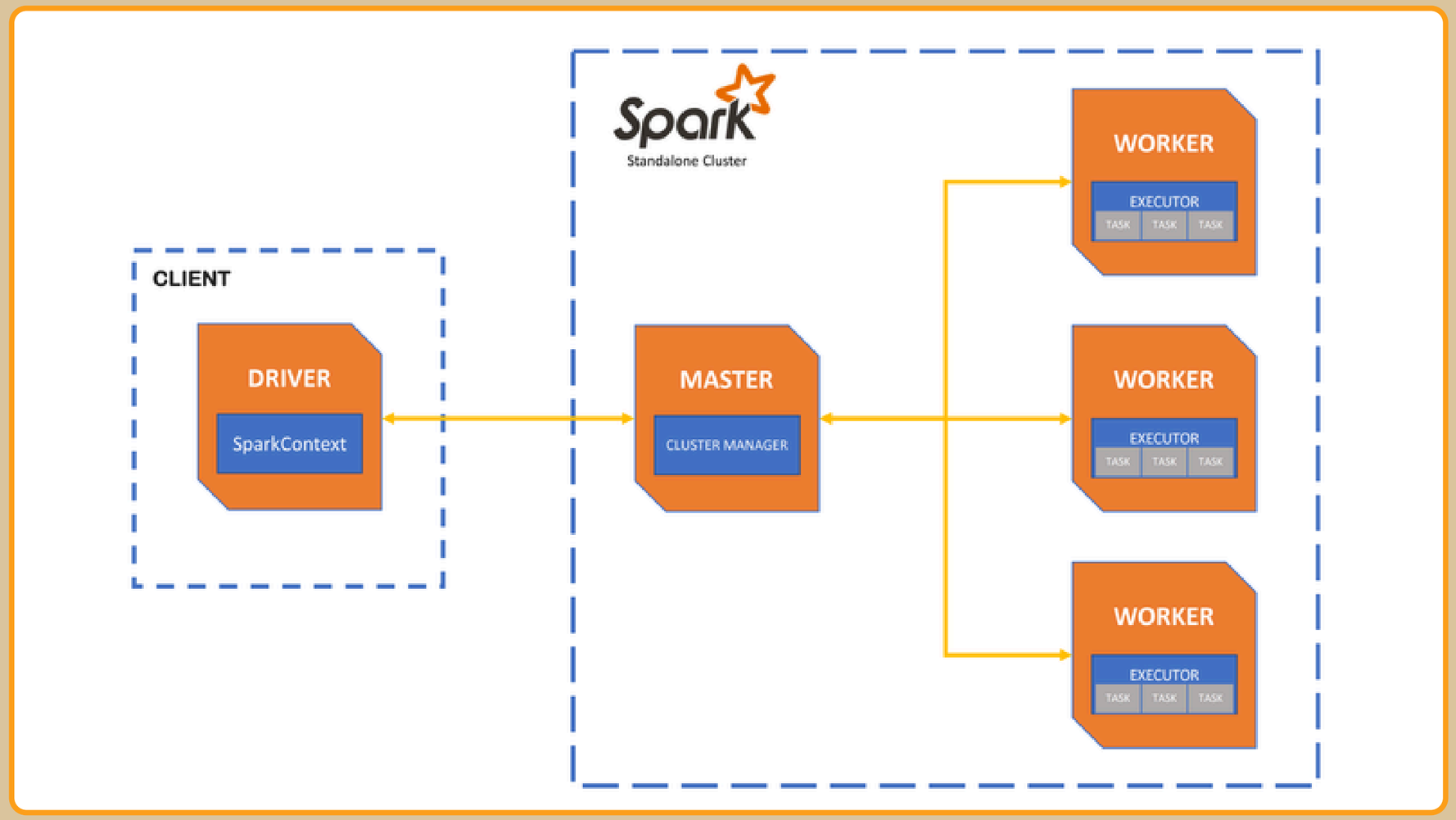
O QUE É O APACHE SPARK?

Um sistema unificado para análise de dados em larga escala de código aberto, ele consegue lidar com lotes, cargas de trabalho de análise e processamento de dados em tempo real.

- **Velocidade**
- **Processamento de stream em tempo real**
- **Análise avançada**

WWW.SPARK.APACHE.ORG

ARQUITETURA APACHE SPARK



AULA 01

TIPOS DE DEPLOY APACHE SPARK

LOCAL

Executa em um único JVM, Laptop ou nó único, executor e gerenciador de cluster Executa no mesmo host.

STANDALONE

Executa em qualquer nó no cluster. Cada Nó Ativa a Própria JVM do Executor. Alocado arbitrariamente para qualquer host

YARN

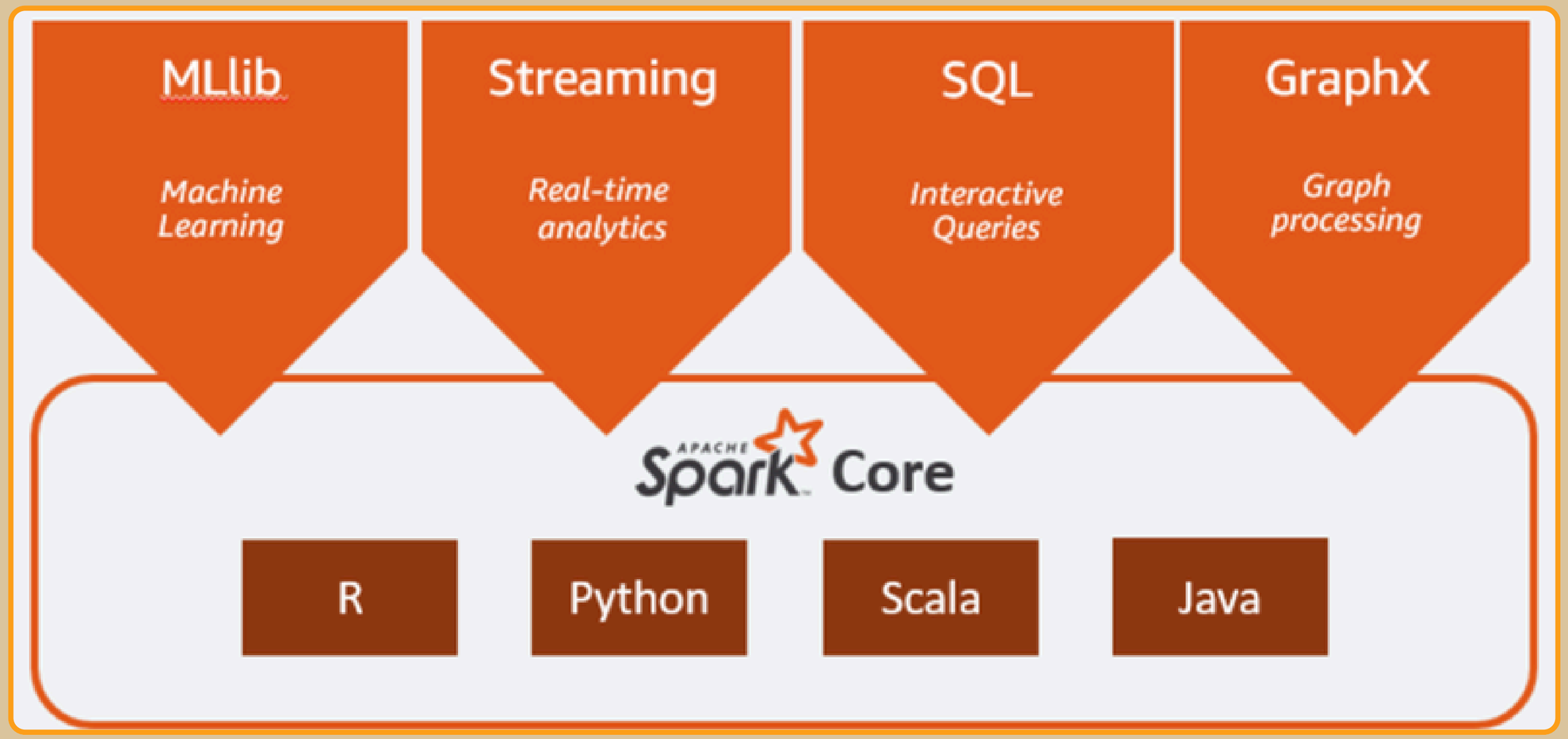
Executa com o mestre de aplicativos YARN (cluster). Gerenciador de nós e contêineres de alocação de recursos para execução

KUBERNETES

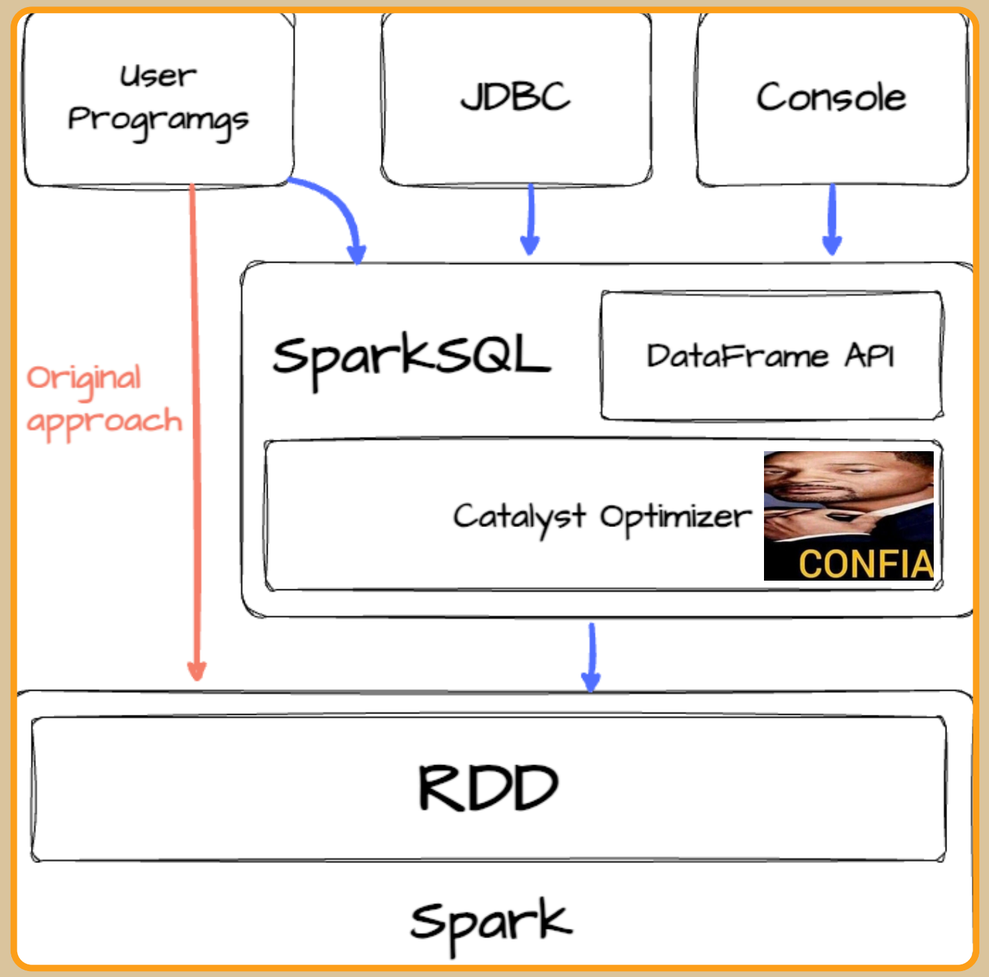
É executado em um pod do Kubernetes. Cada trabalhador é executado no contexto do pod, use o Kubernetes Master para gerenciamento de cluster.

WWW.SPARK.APACHE.ORG

CORE APIS APACHE SPARK



CATALYST OPTIMIZER



AULA 01

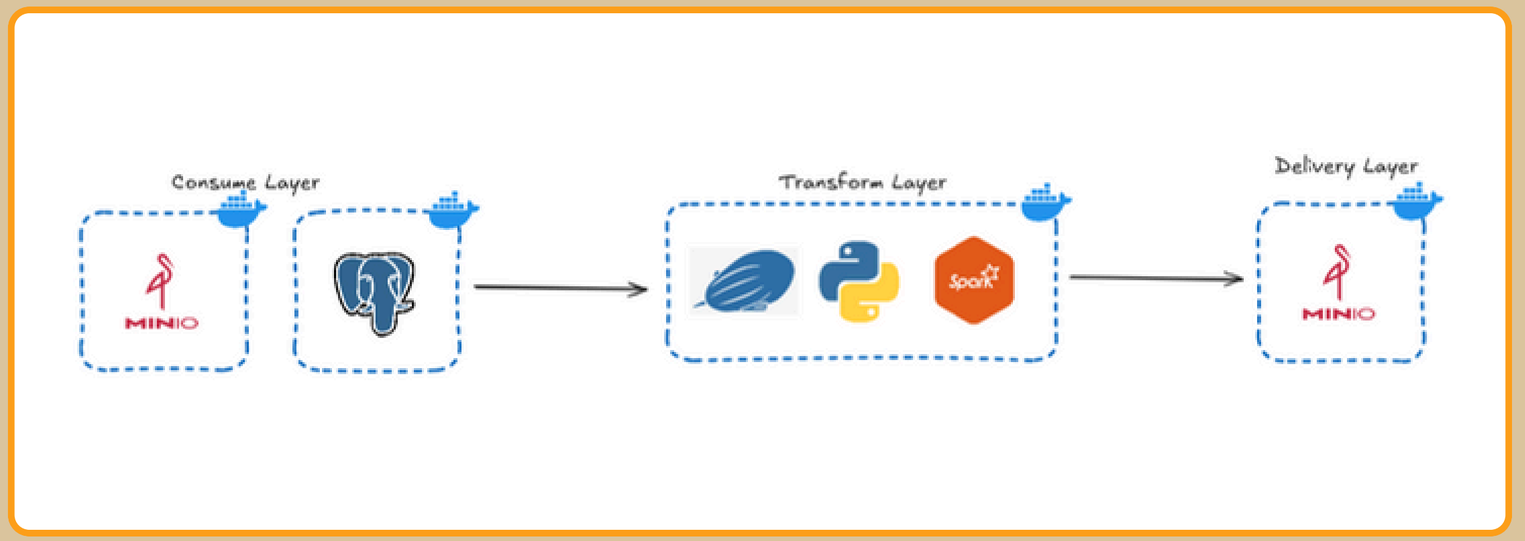
PYSPARK

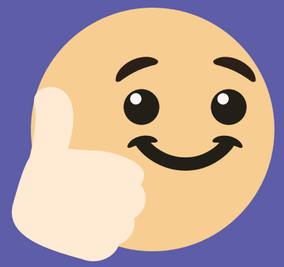
PySpark é a API Python para Apache Spark. Ela permite que você execute processamento de dados em larga escala e em tempo real em um ambiente distribuído usando Python.

WWW.SPARK.APACHE.ORG

AULA 01

AULA PRÁTICA





OBRIGADO !