

# Streaming de Dados em Tempo Real: Aula 1

*Prof. Felipe Timbó*



# Objetivos

Estudar/investigar princípios, técnicas e ferramentas necessárias para lidar com **streaming de dados**.

Desenvolver soluções em tempo real, isto é, à medida que dados são produzidos.

Resolver problemas relacionados a **streaming de dados em tempo real** com Kafka e Spark

# Ementa

- **Dia 1:**
  - Introdução a Streaming de Dados
  - Introdução ao Apache Kafka
  - Setup do ambiente
  
- **Dia 2:**
  - Data Ingestion com Apache Kafka
  - Kafka Web Project

# Ementa

- Dia 3:
  - Introdução ao Apache Spark
- Dia 4:
  - Processamento de dados de Streaming com Apache Spark

# Tecnologias e Ferramentas deste Curso

Máquina Virtual (VirtualBox)

Linux (Ubuntu)

Python

VS Code

Apache Kafka

Apache Spark

# Metodologia

Aulas expositivas com discussões

Práticas remotas

Leituras

Tarefas individuais

Projeto final (até 3 pessoas)

# Recursos

Lista de e-mails:

[uni7-ciencia-de-dados-turma10@googlegroups.com](mailto:uni7-ciencia-de-dados-turma10@googlegroups.com)

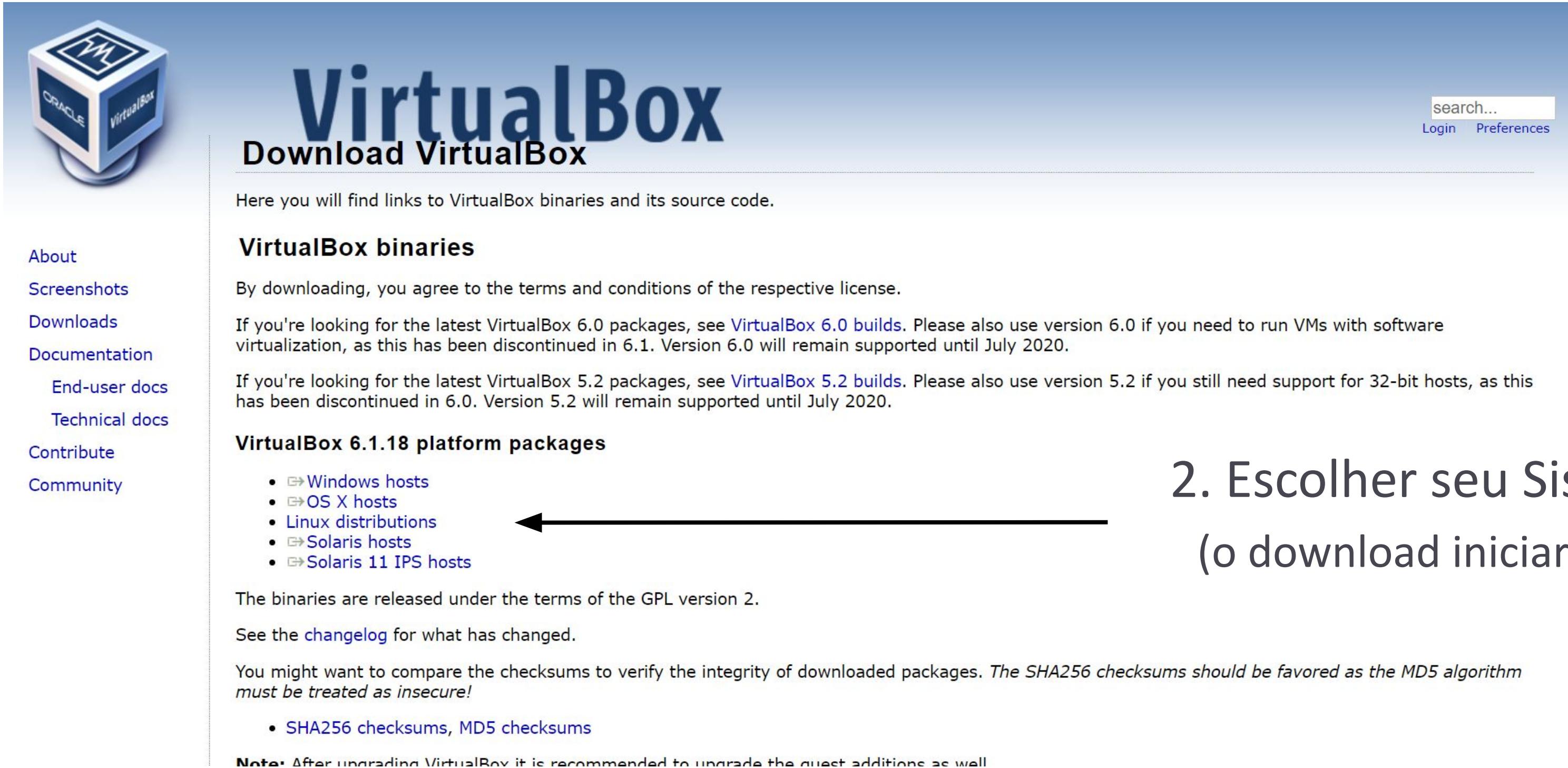
Arquivos (slides, livros e artigos)

Github: <https://github.com/felipetimbo/streaming-data-course>

Antes de tudo...

# Download do VirtualBox

1. Acessar <https://www.virtualbox.org/wiki/Downloads>



The screenshot shows the 'VirtualBox Download' page. At the top left is the Oracle VM VirtualBox logo. The main title 'VirtualBox' is in large blue letters, with 'Download VirtualBox' below it. To the right are 'search...', 'Login', and 'Preferences' buttons. On the left, a sidebar lists links: 'About', 'Screenshots', 'Downloads', 'Documentation', 'End-user docs', 'Technical docs', 'Contribute', and 'Community'. The main content area starts with a note about finding binaries and source code. It then has two sections: 'VirtualBox binaries' and 'VirtualBox 6.1.18 platform packages'. The 'VirtualBox binaries' section contains text about license terms and links to 'VirtualBox 6.0 builds' and 'VirtualBox 5.2 builds'. The 'VirtualBox 6.1.18 platform packages' section lists platforms: Windows hosts, OS X hosts, Linux distributions, Solaris hosts, and Solaris 11 IPS hosts. An arrow points from the 'VirtualBox 6.1.18 platform packages' section to the second step of the guide.

Here you will find links to VirtualBox binaries and its source code.

### VirtualBox binaries

By downloading, you agree to the terms and conditions of the respective license.

If you're looking for the latest VirtualBox 6.0 packages, see [VirtualBox 6.0 builds](#). Please also use version 6.0 if you need to run VMs with software virtualization, as this has been discontinued in 6.1. Version 6.0 will remain supported until July 2020.

If you're looking for the latest VirtualBox 5.2 packages, see [VirtualBox 5.2 builds](#). Please also use version 5.2 if you still need support for 32-bit hosts, as this has been discontinued in 6.0. Version 5.2 will remain supported until July 2020.

### VirtualBox 6.1.18 platform packages

- [Windows hosts](#)
- [OS X hosts](#)
- [Linux distributions](#)
- [Solaris hosts](#)
- [Solaris 11 IPS hosts](#)

The binaries are released under the terms of the GPL version 2.

See the [changelog](#) for what has changed.

You might want to compare the checksums to verify the integrity of downloaded packages. *The SHA256 checksums should be favored as the MD5 algorithm must be treated as insecure!*

- [SHA256 checksums](#), [MD5 checksums](#)

Note: After upgrading VirtualBox it is recommended to upgrade the guest additions as well

2. Escolher seu Sistema Operacional  
(o download iniciará automaticamente)

# Download do Ubuntu

3. Acessar: <https://ubuntu.com/download/desktop>

4. Realizar o download do  
Ubuntu 22.04.2 LTS

## Ubuntu 22.04.2 LTS

The latest LTS version of Ubuntu, for desktop PCs and laptops. LTS stands for long-term support — which means five years of free security and maintenance updates, guaranteed until April 2027.

[Ubuntu 22.04 LTS release notes](#)

Recommended system requirements:

- ✓ 2 GHz dual-core processor or better
- ✓ 4 GB system memory
- ✓ 25 GB of free hard drive space
- ✓ Internet access is helpful
- ✓ Either a DVD drive or a USB port for the installer media

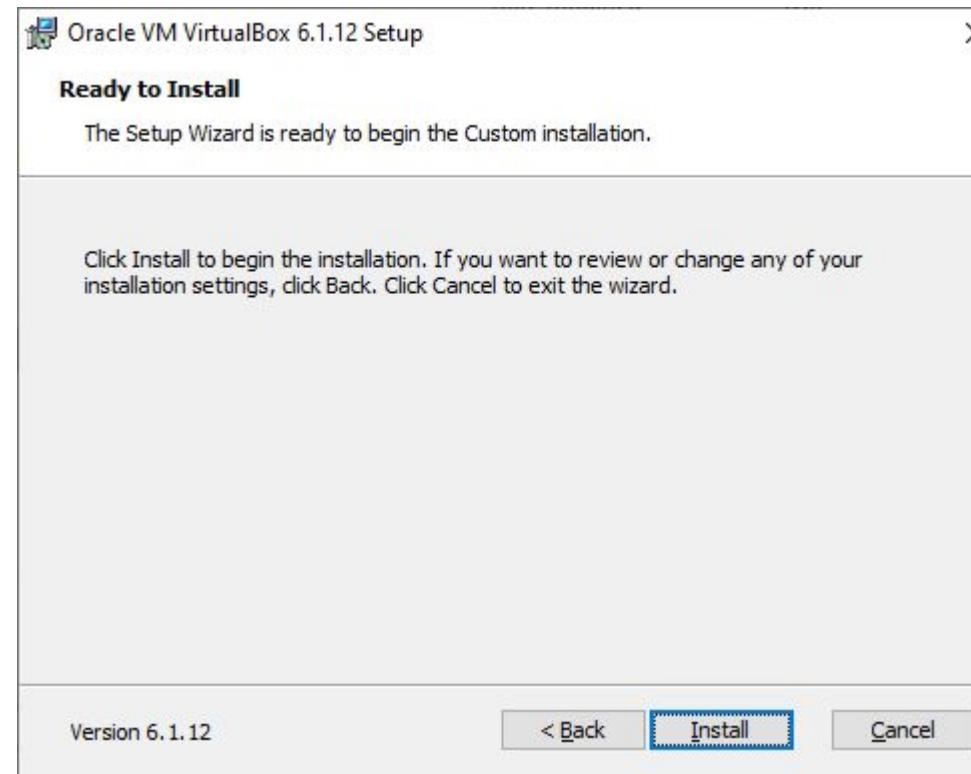
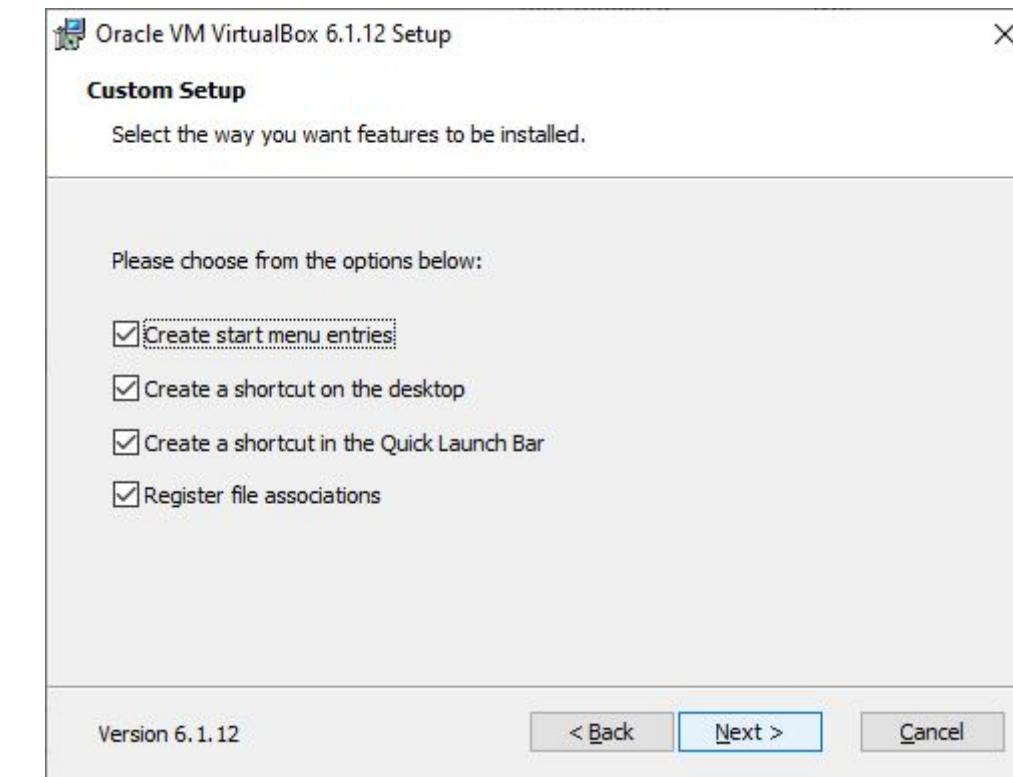
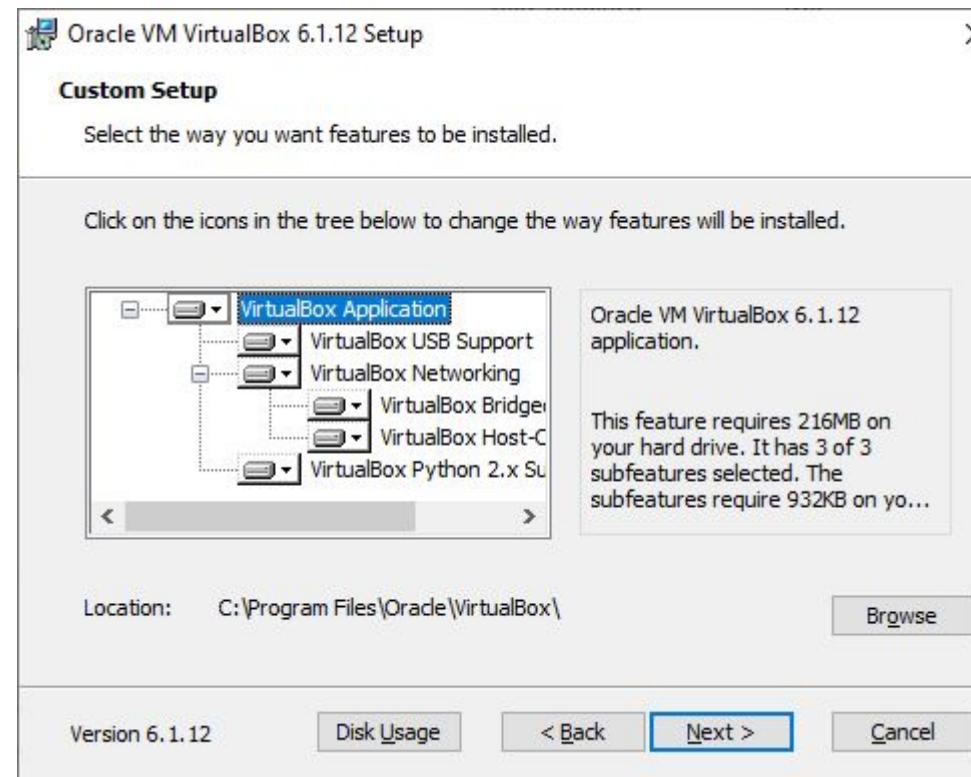


[Download](#)

For other versions of Ubuntu Desktop including torrents, the network installer, a list of local mirrors and past releases [see our alternative downloads](#).

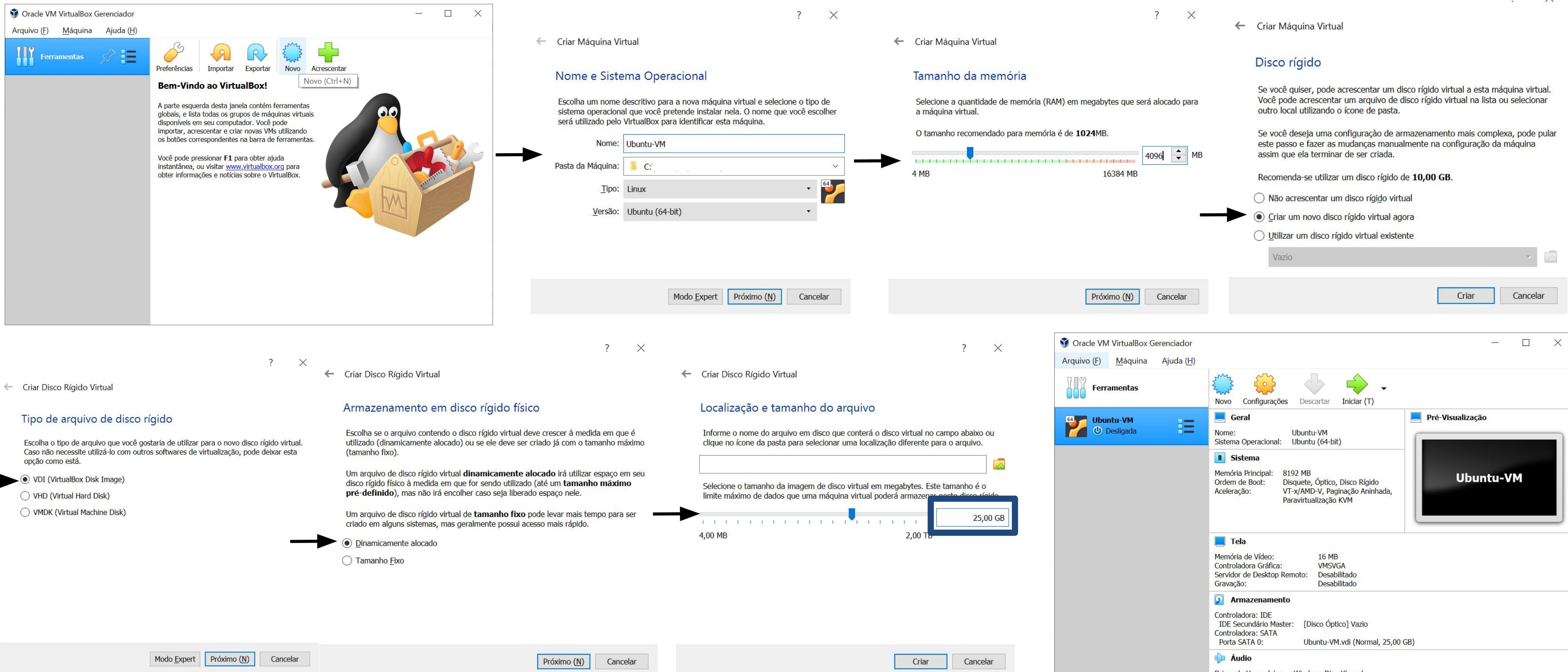
# Instalação do VirtualBox

## 5. Instalar o VirtualBox



# Criação de uma VM (Virtual Machine)

## 6. Criar uma nova Máquina Virtual Ubuntu



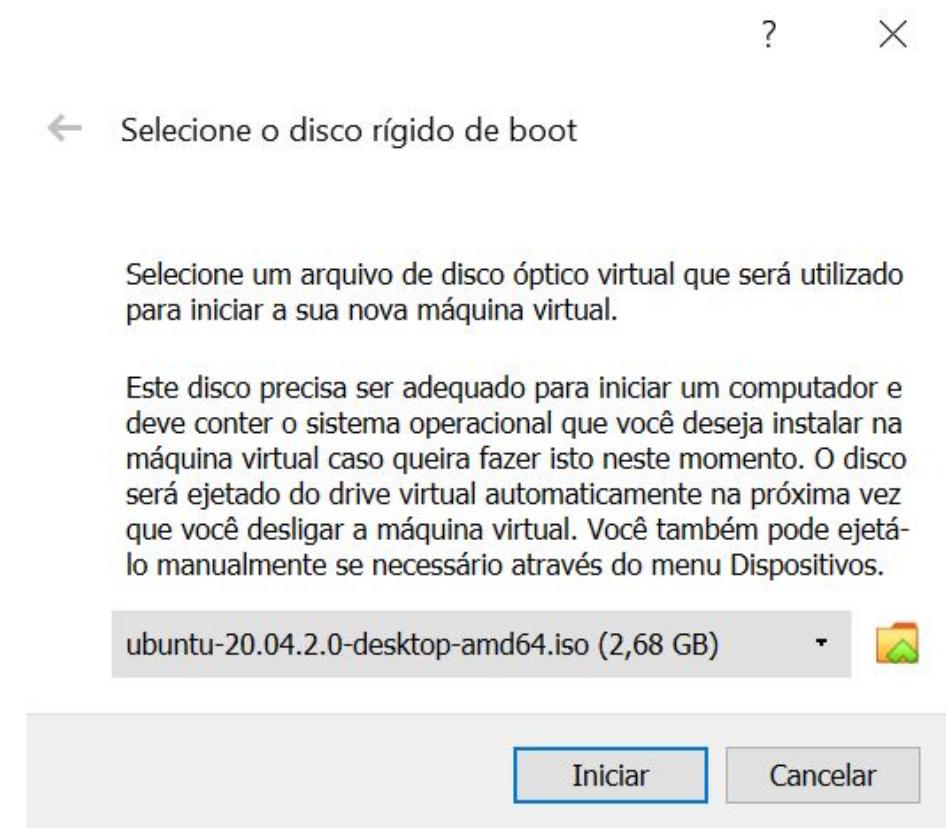
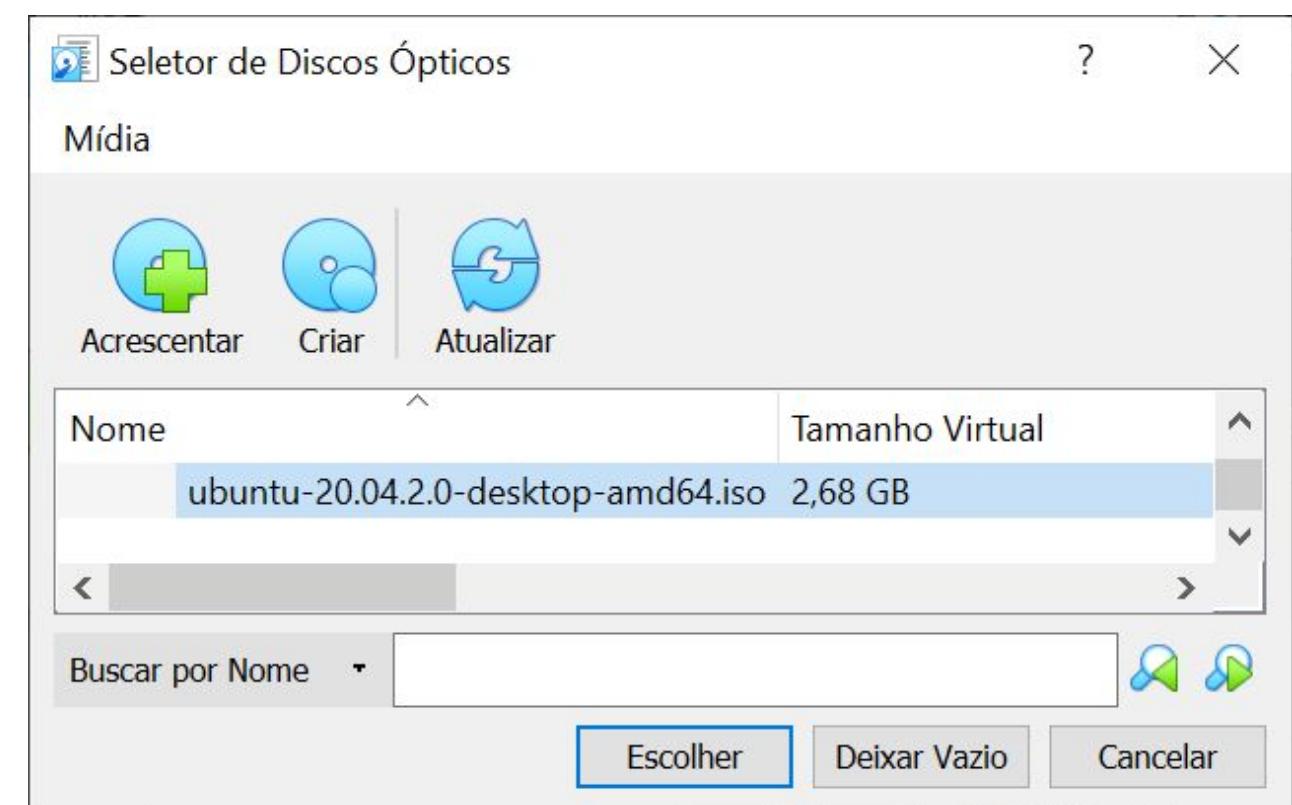
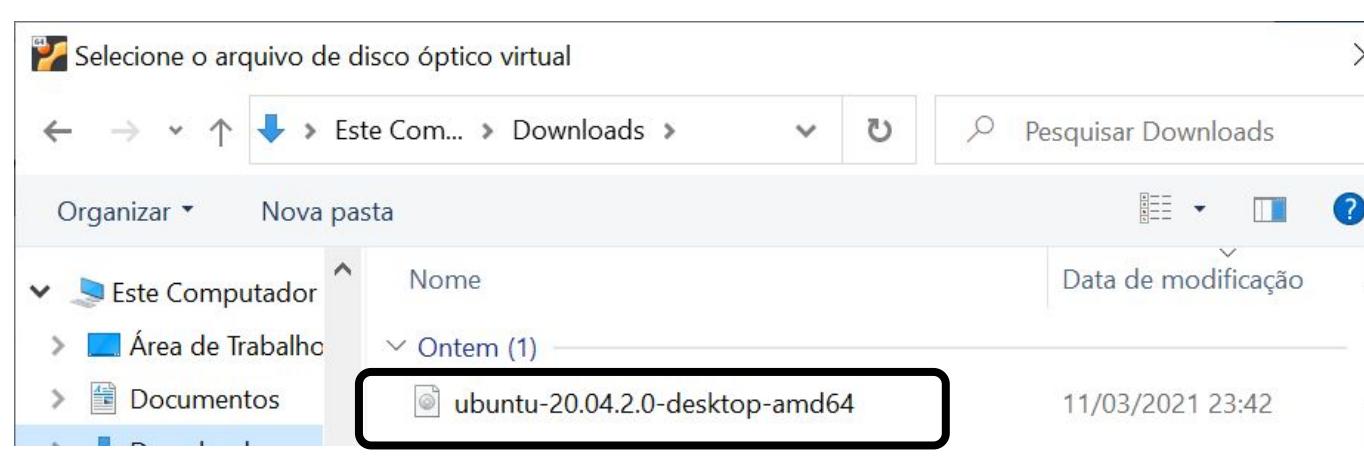
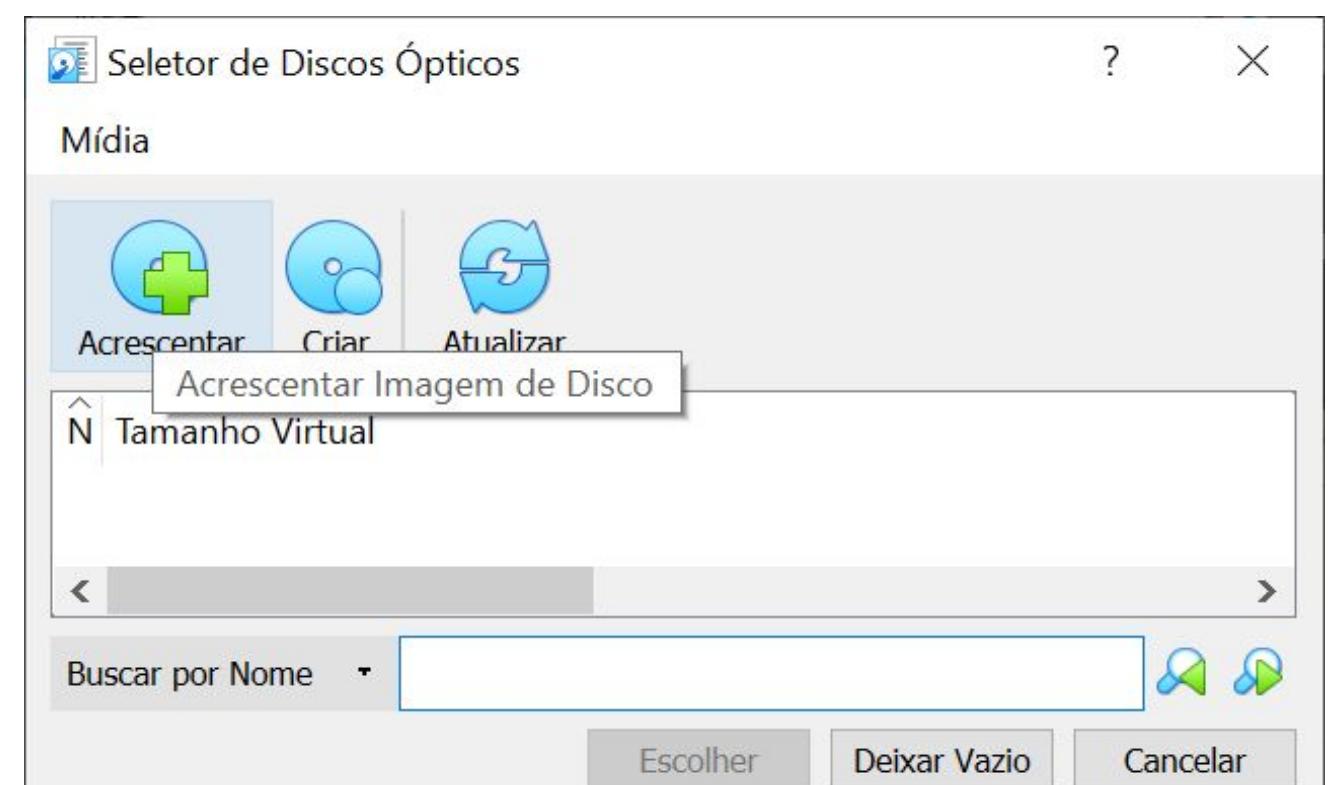
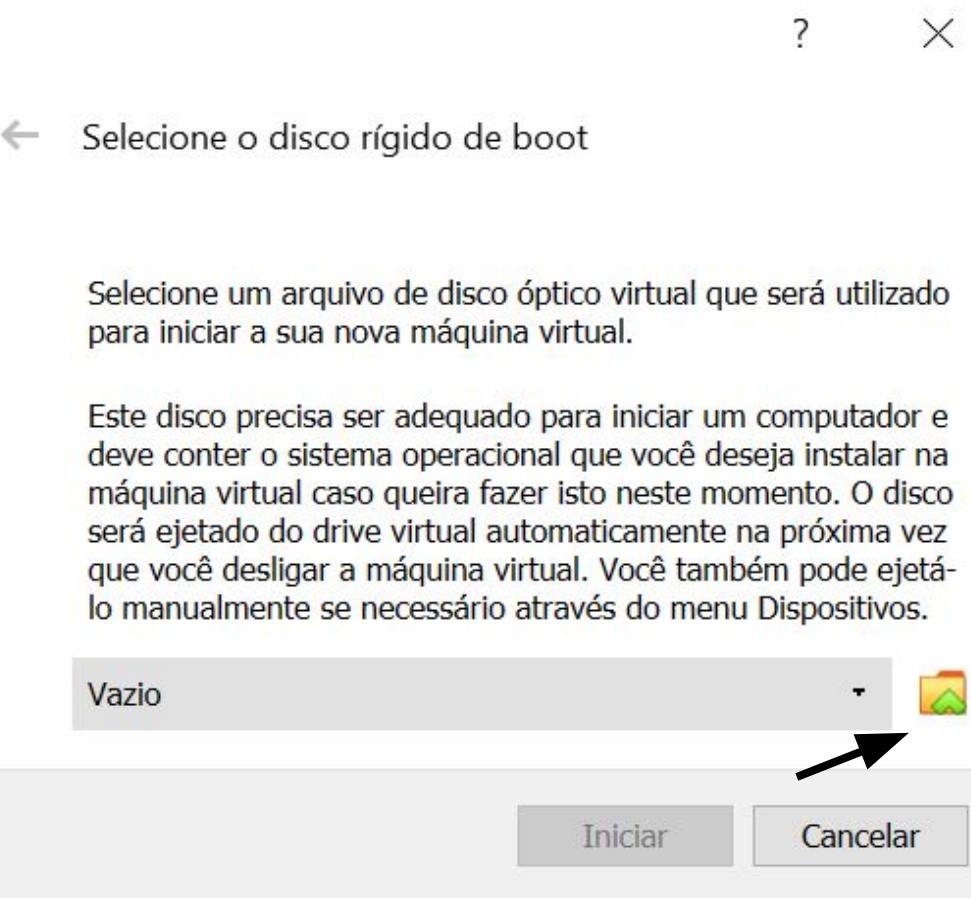
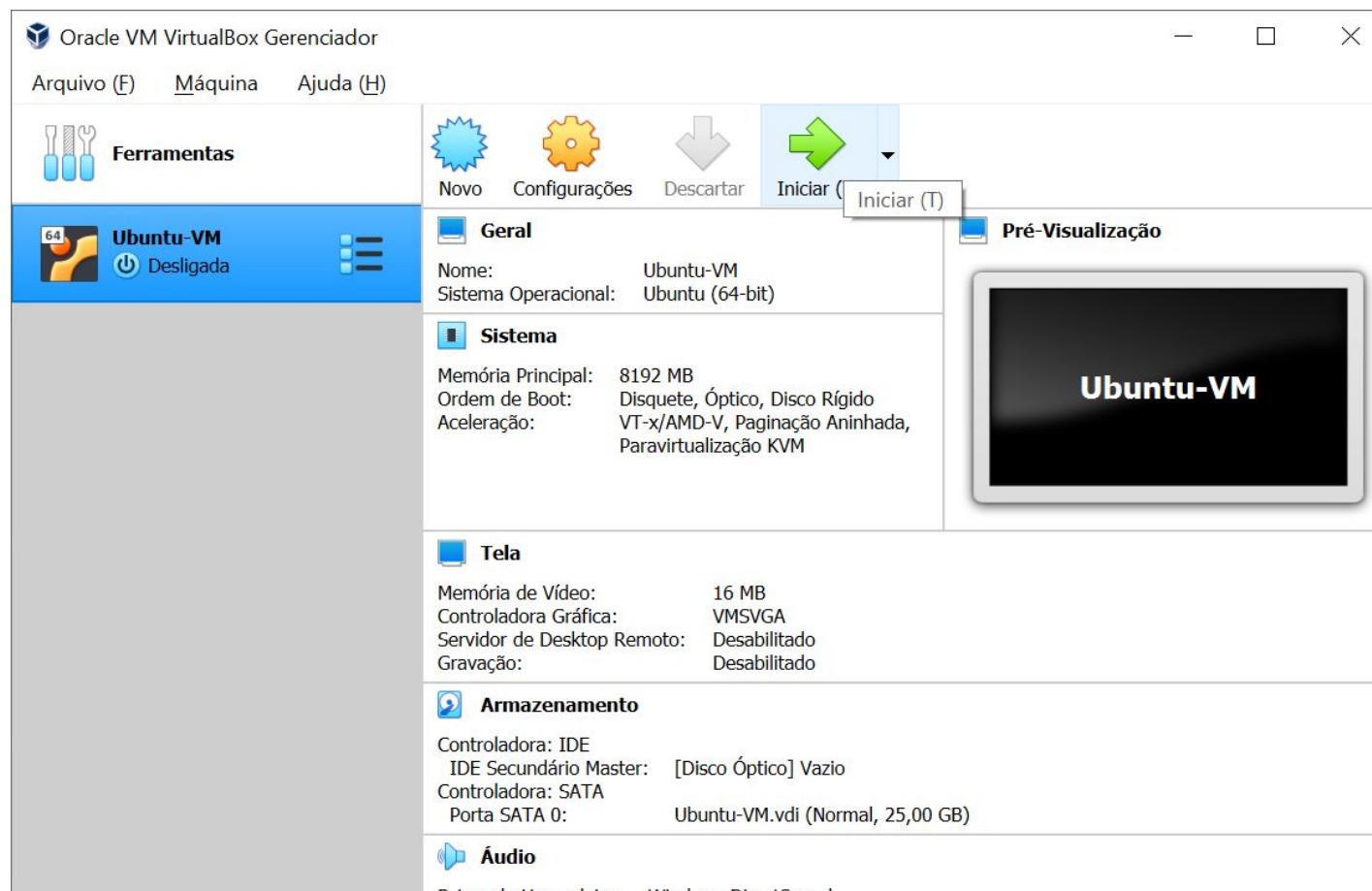
The image shows a step-by-step guide for creating a new Virtual Machine (VM) in Oracle VM VirtualBox. It consists of five windows arranged horizontally:

- Oracle VM VirtualBox Gerenciador:** Main interface showing the 'Bem-Vindo ao VirtualBox!' (Welcome to VirtualBox!) screen.
- Criar Máquina Virtual - Nome e Sistema Operacional:** Step 1 of the wizard. Name: Ubuntu-VM, Type: Linux, Version: Ubuntu (64-bit).
- Criar Máquina Virtual - Tamanho da memória:** Step 2 of the wizard. RAM size: 4096 MB (1024 MB recommended).
- Criar Máquina Virtual - Disco rígido:** Step 3 of the wizard. Option selected: Criar um novo disco rígido virtual agora.
- Criar Disco Rígido Virtual - Tipo de arquivo de disco rígido:** Step 4 of the wizard. Type selected: VDI (VirtualBox Disk Image).
- Criar Disco Rígido Virtual - Armazenamento em disco rígido físico:** Step 5 of the wizard. Allocation type: Dinamicamente alocado (Dynamic allocation).
- Criar Disco Rígido Virtual - Localização e tamanho do arquivo:** Step 6 of the wizard. File size: 25,00 GB.
- Oracle VM VirtualBox Gerenciador:** Final view of the VirtualBox Manager showing the newly created VM 'Ubuntu-VM' in the list.

Arrows indicate the flow from one step to the next, and a large arrow points from the final configuration window back to the main manager window.

# Criação de uma VM (Virtual Machine)

## 7. Iniciar a Máquina Virtual Ubuntu



# Instalação do Ubuntu

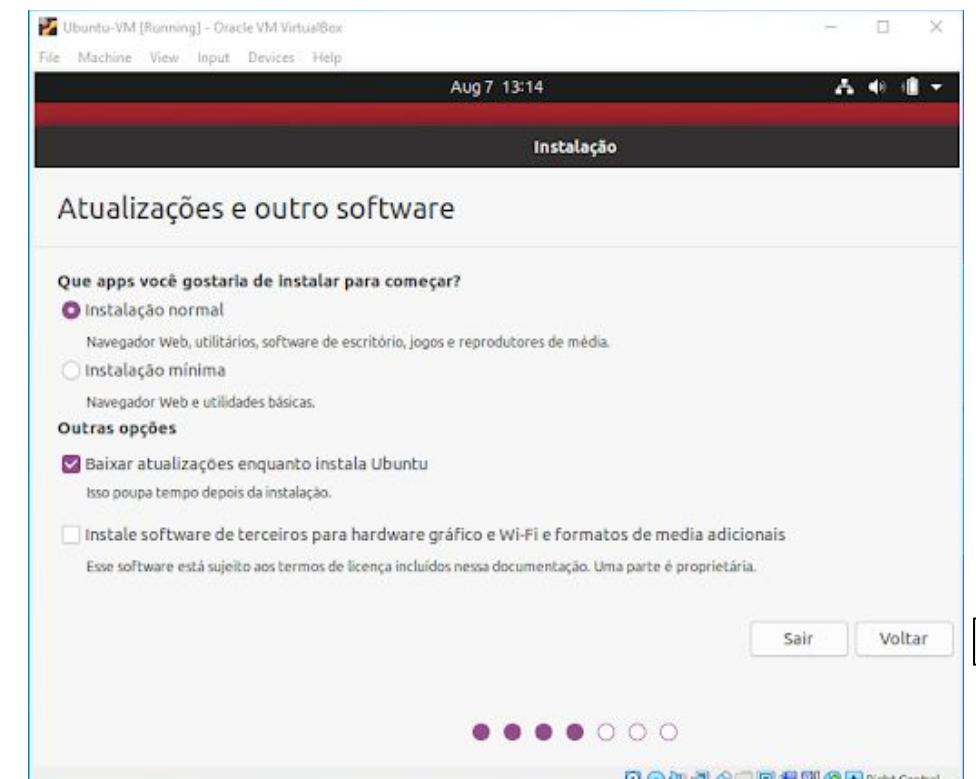
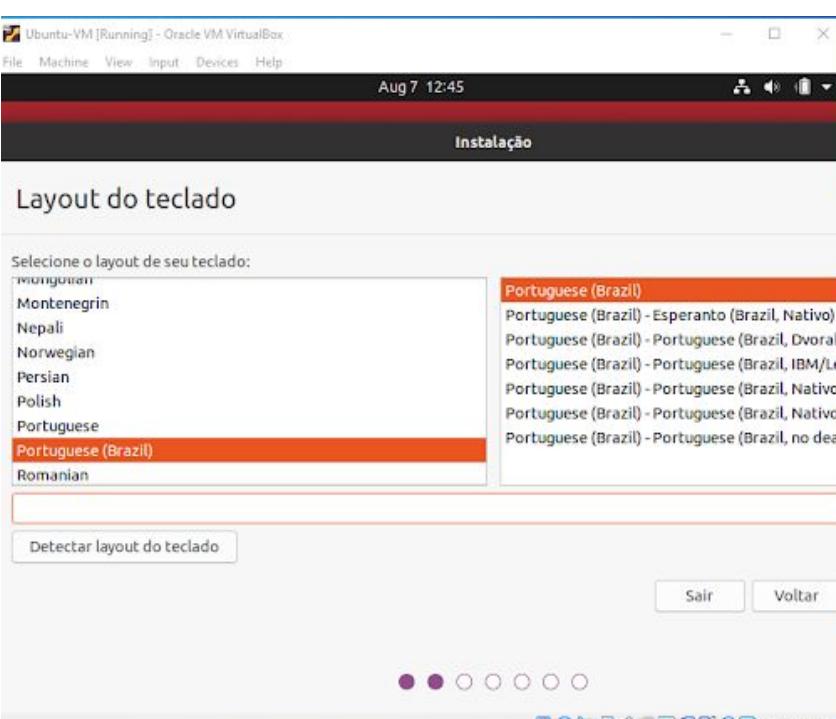
## 8. Instalar o Ubuntu na VM



dois cliques



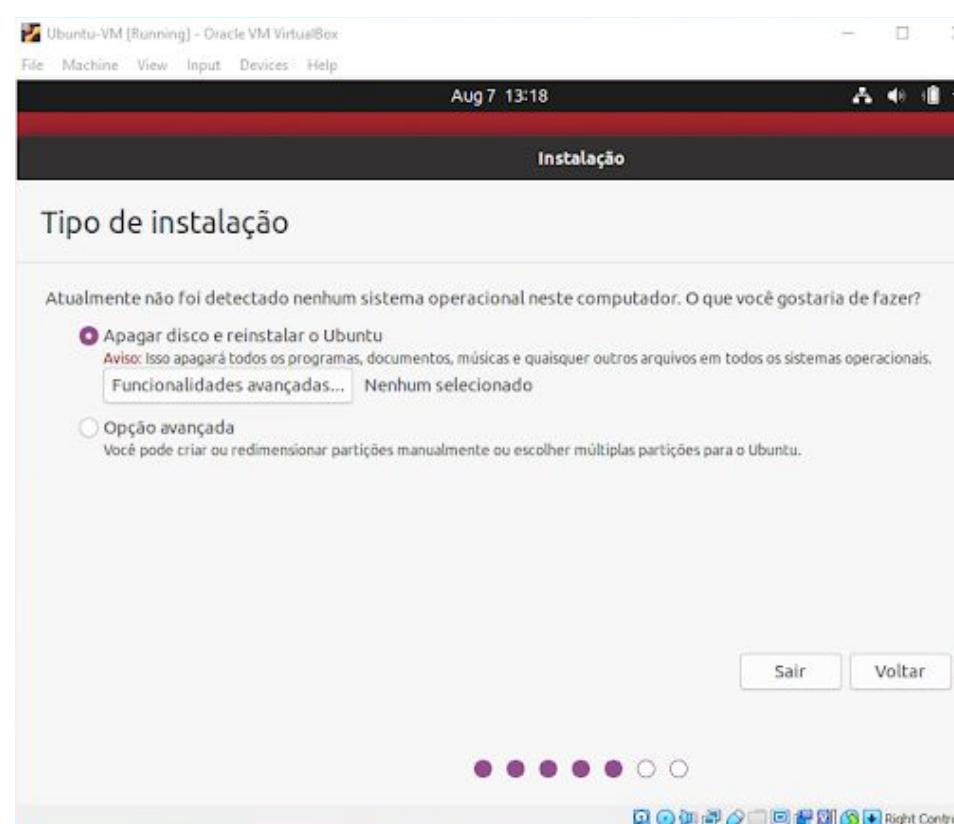
→



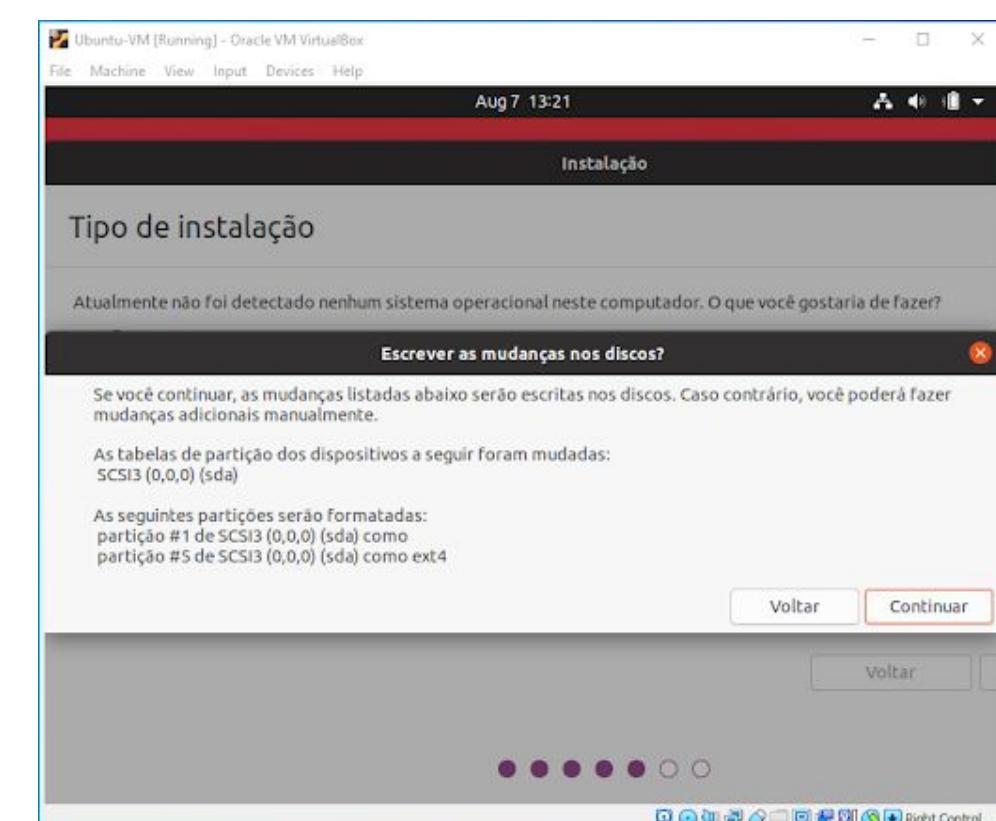
continuar



o botão 'continuar' pode  
estar escondido aqui.  
Acessá-lo via tecla 'tab'

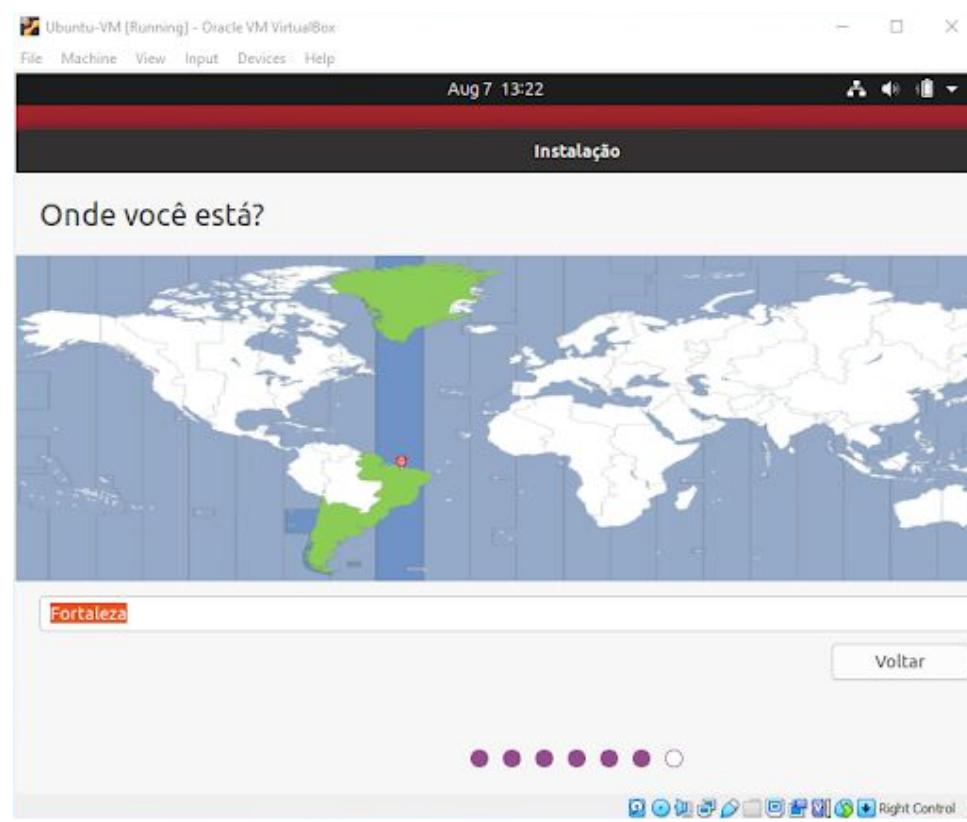


o botão 'continuar' pode  
estar escondido aqui.  
Acessá-lo via tecla 'tab'

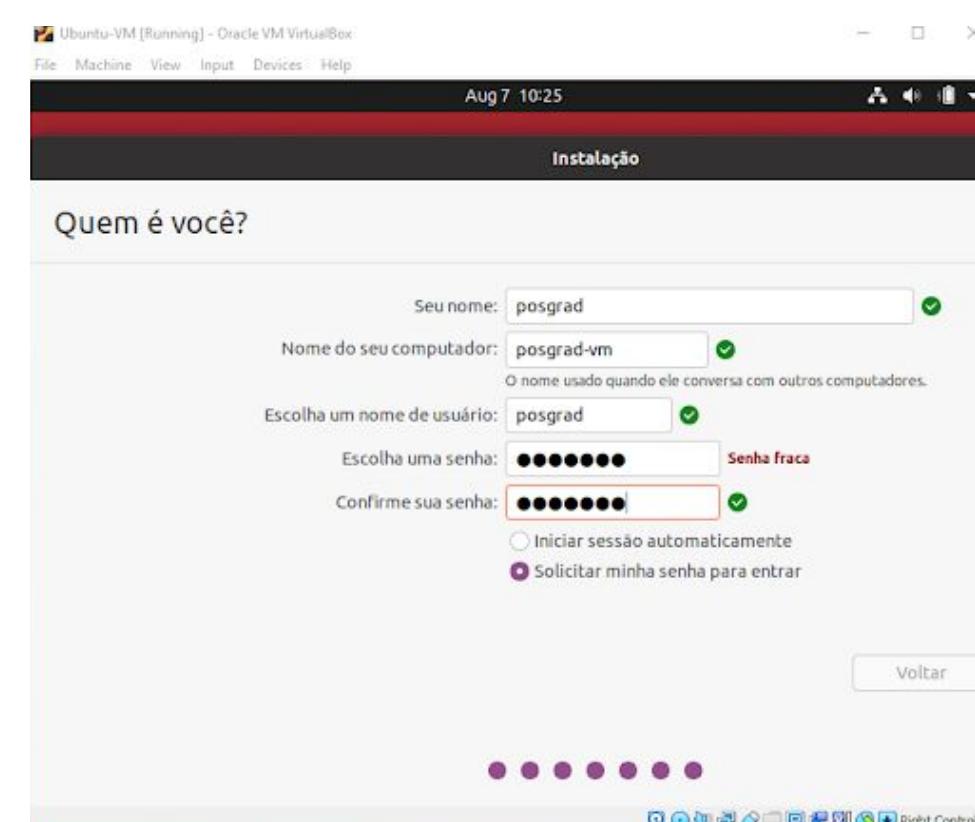


# Instalação do Ubuntu

## 8. Instalar o Ubuntu na VM

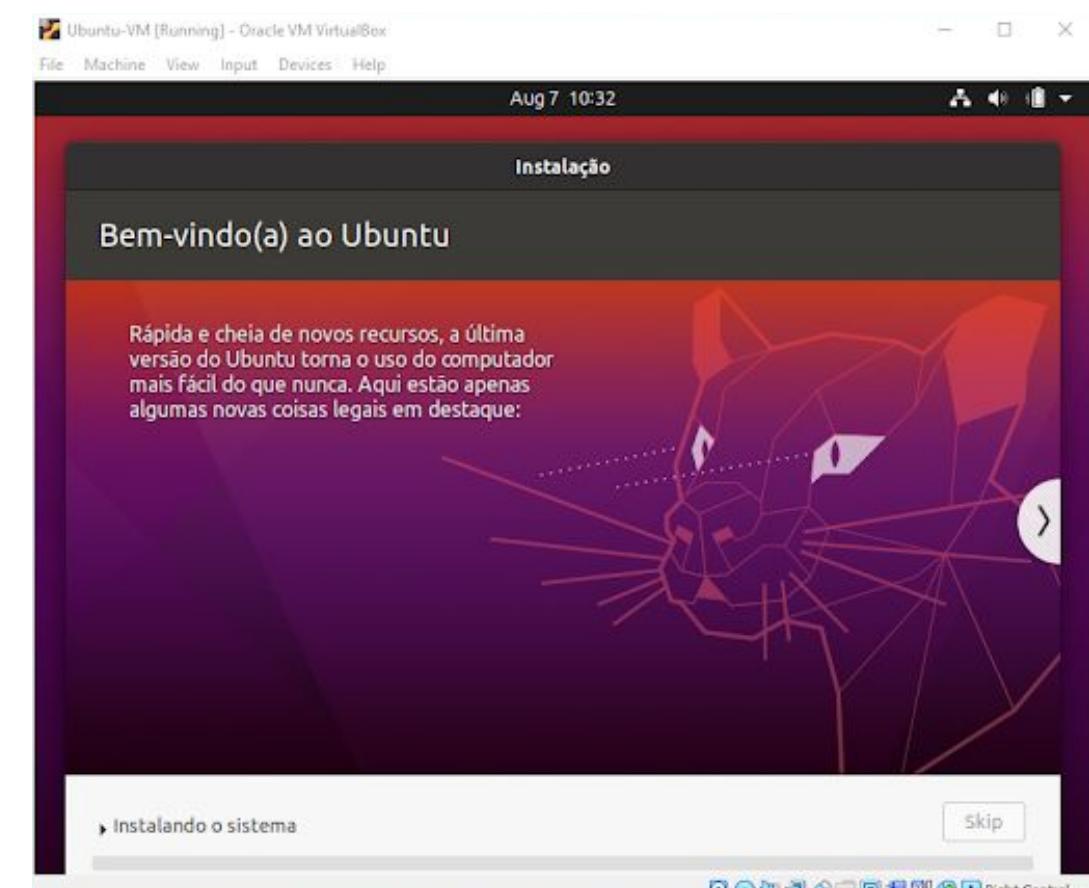


o botão 'continuar' pode  
estar escondido aqui.  
Acessá-lo via tecla 'tab'



o botão 'continuar' pode  
estar escondido aqui.  
Acessá-lo via tecla 'tab'

{ usuário: posgrad  
senha: posgrad



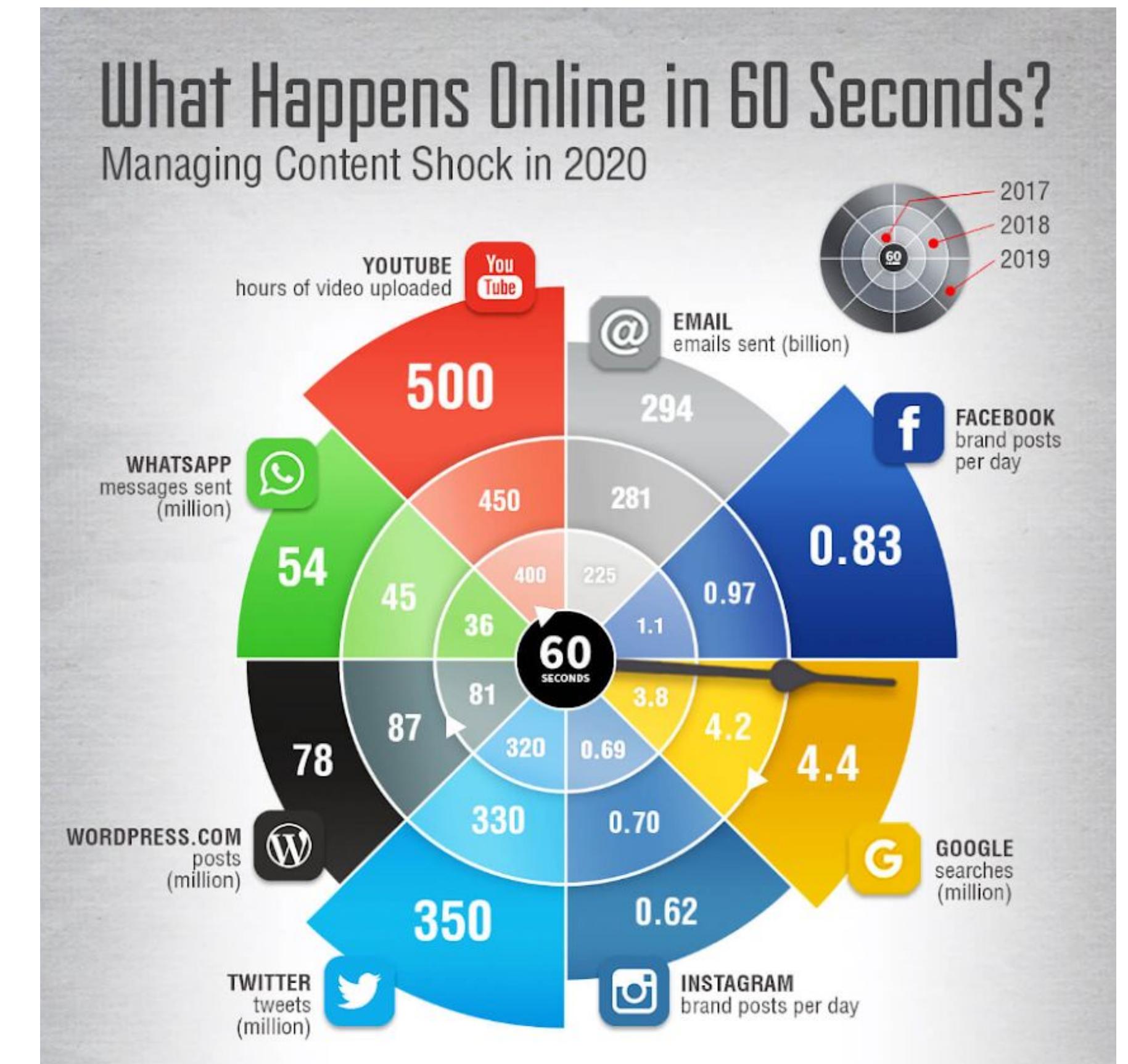
Deixar instalando...

# Introdução

# O que acontece em 60 segundos?

Upload de 500h de vídeo  
no YouTube

54 milhões de  
mensagens no Whatsapp



# Big Data

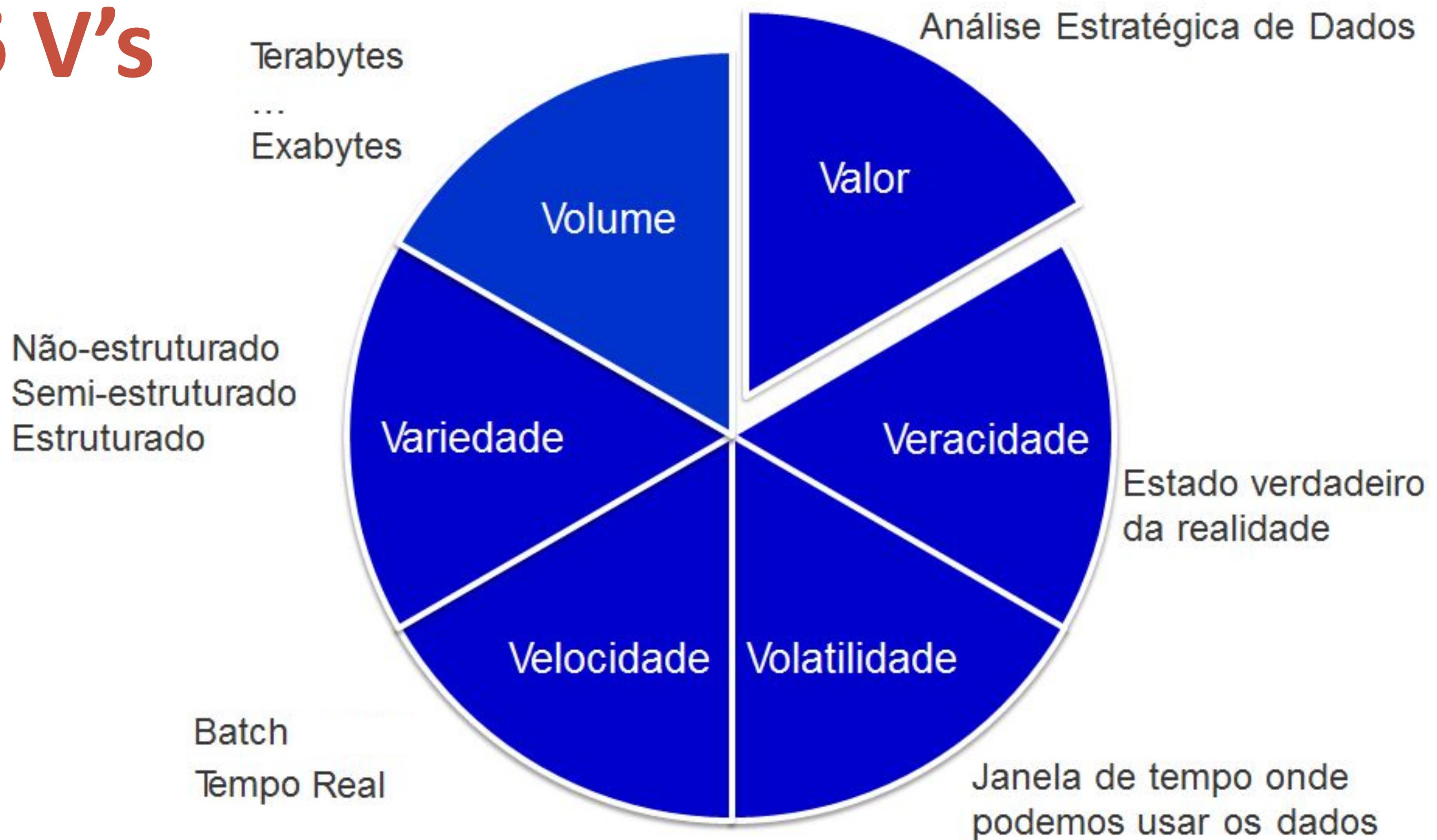
Big Data são dados que excedem o **armazenamento, o processamento e a capacidade dos sistemas convencionais**

Volume de dados muito grande

Dados são gerados rapidamente

Dados não se encaixam nas estruturas de arquiteturas de sistemas atuais

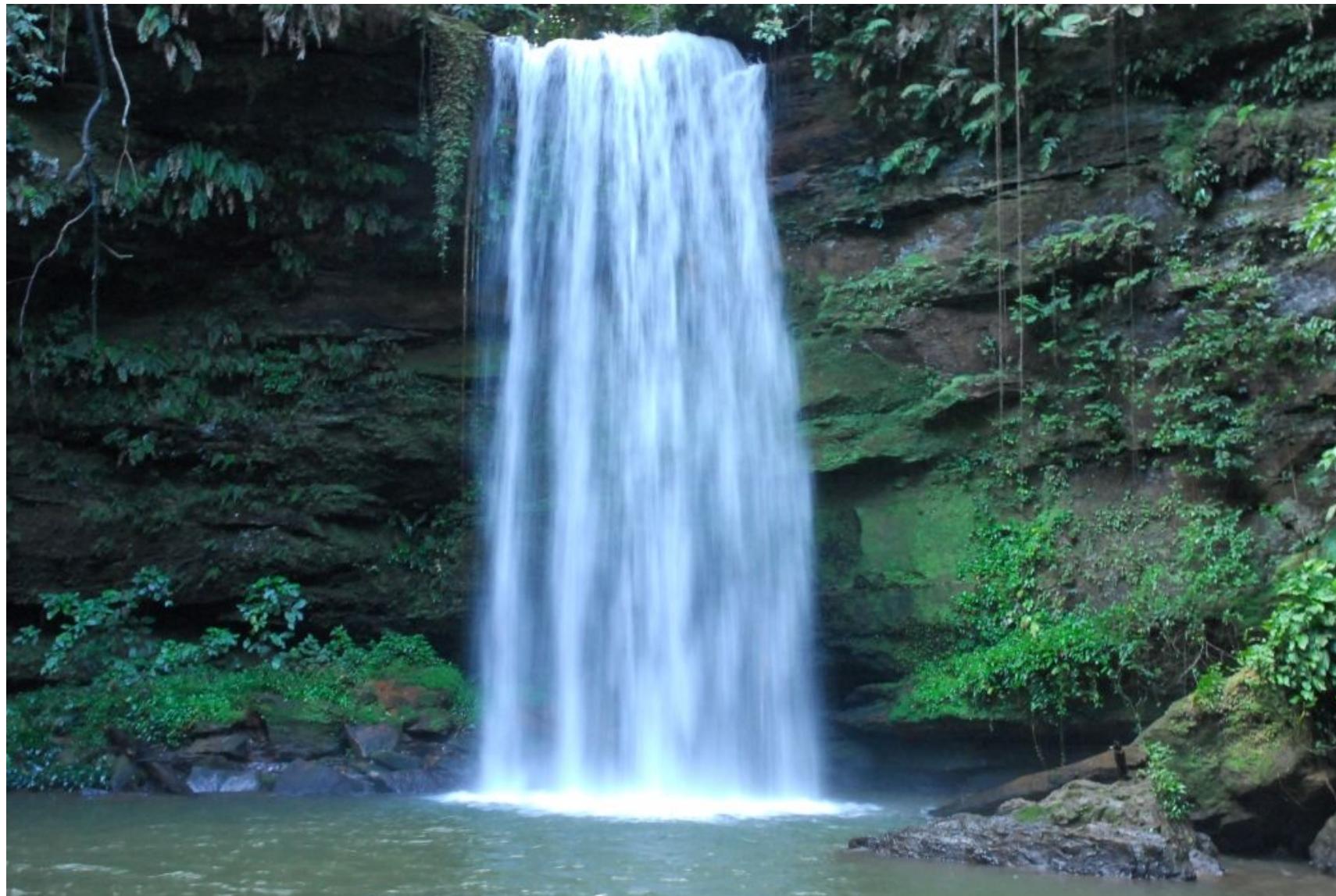
# 6 V's



# Streaming

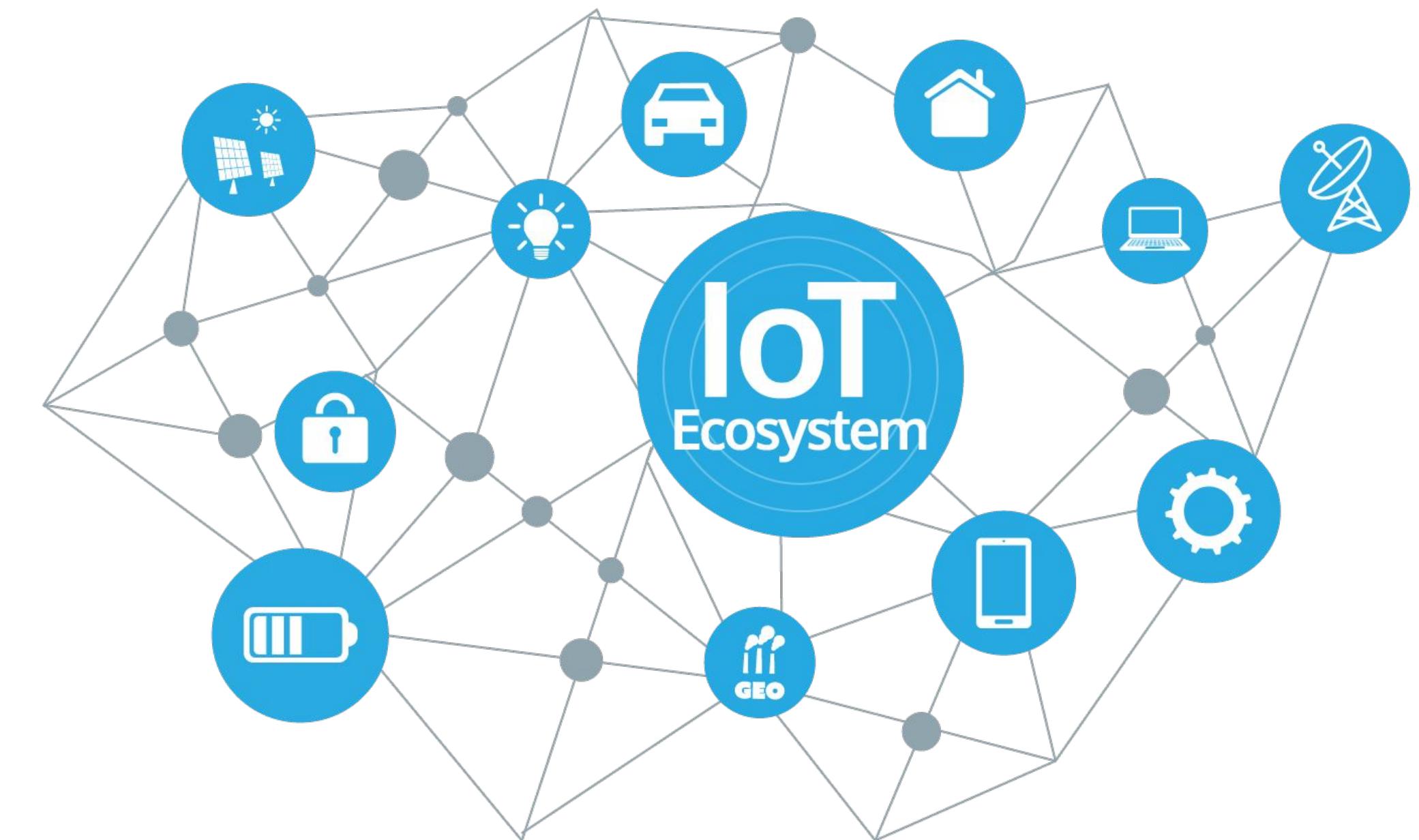
# Streaming

Fluxo contínuo (contínuo  $\neq$  constante).



# Streaming de dados

Fluxo contínuo de dados.



# Streaming de dados: Exemplos

- Sensores (IoT)
- Tráfego de rede
- Registros de call center
- Tendências em redes sociais
- Serviços de áudio e vídeo
- Análise de log
- Estatísticas de sites web



# Tipos de streaming de dados

Dados de texto: web, log

Dados relacionais: tabelas, transações

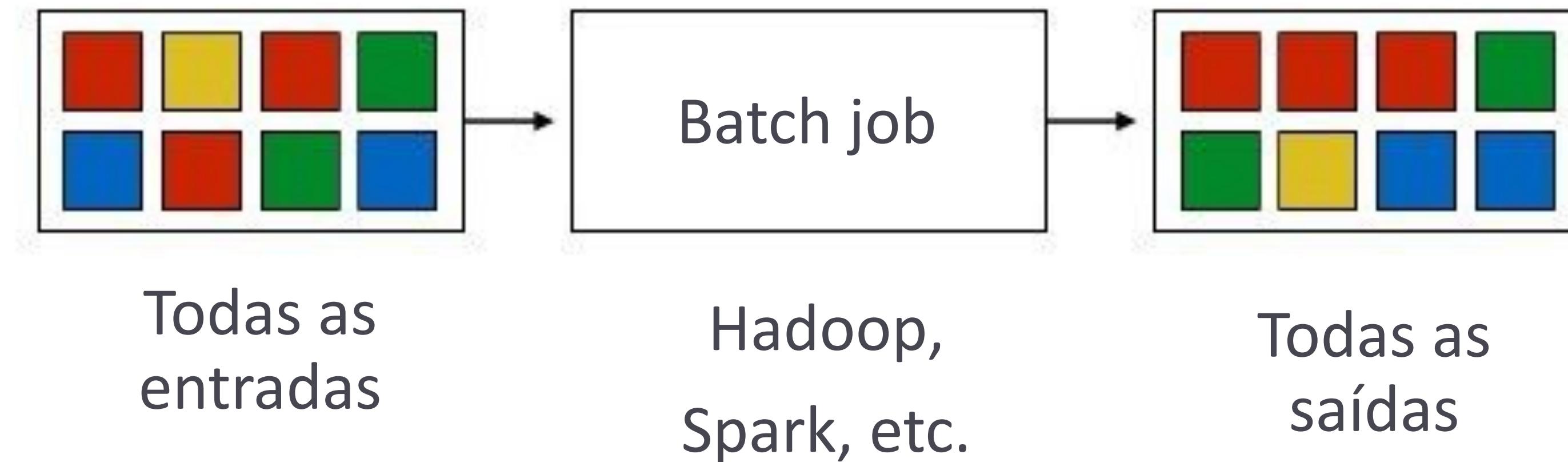
Dados semi-estruturados: XML, json

Dados em grafo: redes sociais

Dados de mobilidade: coordenadas geográficas x tempo

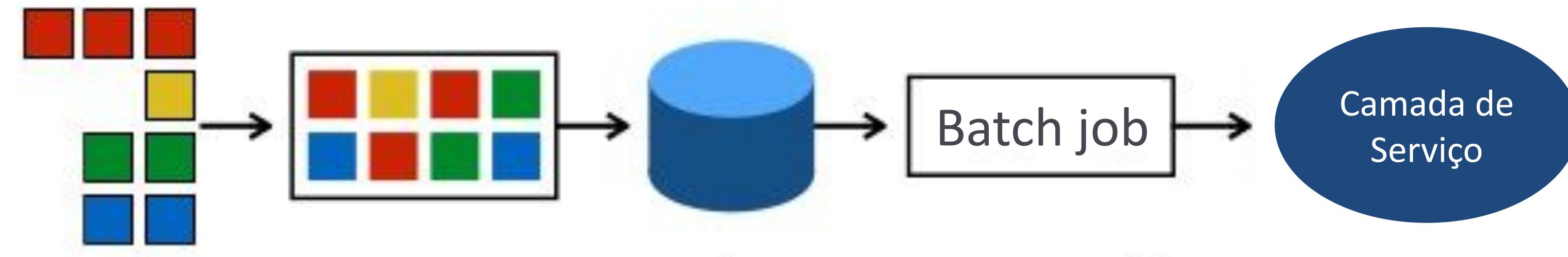
Etc.

# Processamento em Batch



# Processamento de Streaming

Em geral:



continuamente  
produzido

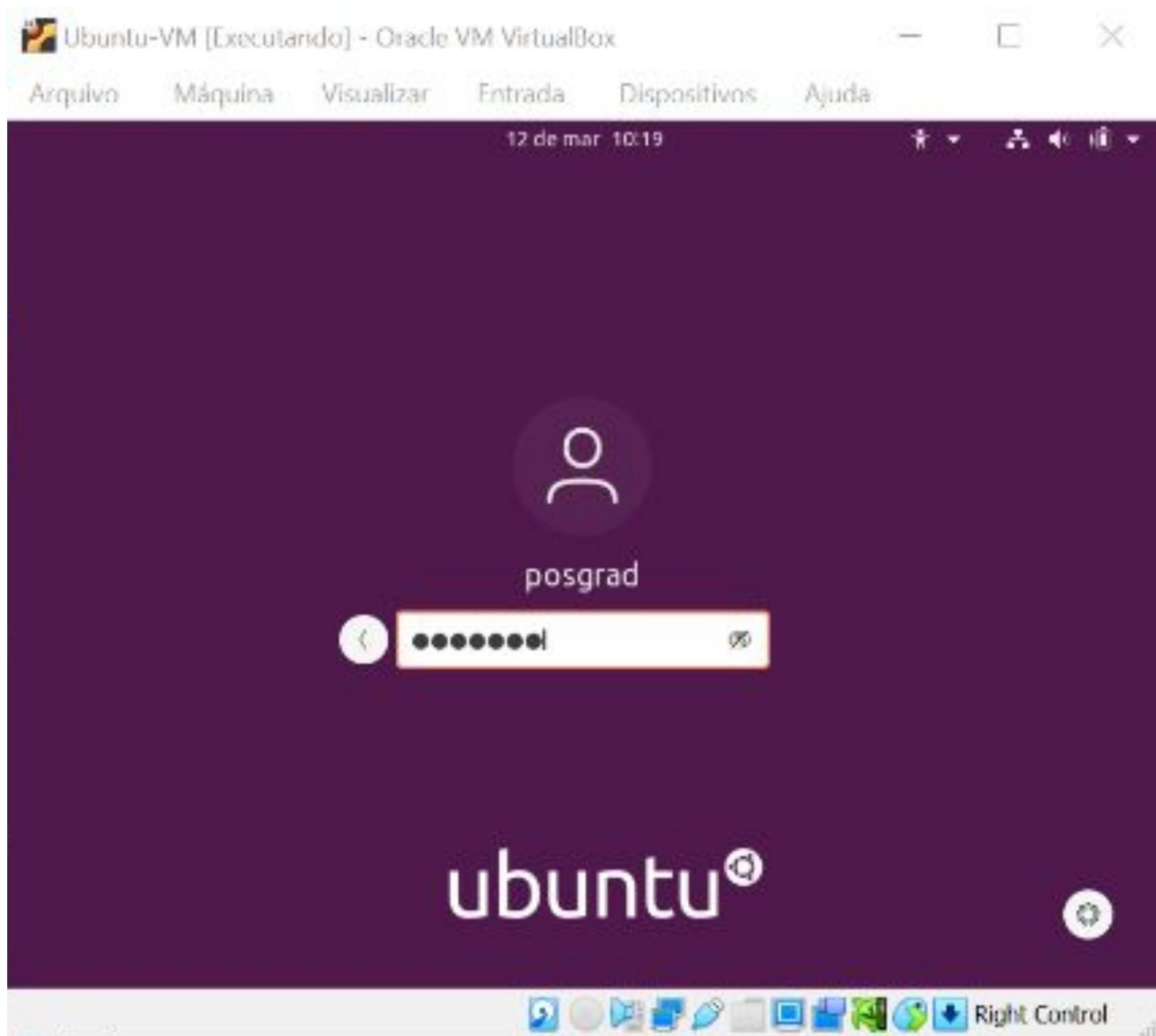
arquivos são  
streams finitos

periodicamente  
executado

Continuando  
nossa Setup...

# Configuração da VM Ubuntu

## 9. Logar na VM



# Atualizando o Ubuntu

10. Abrir o terminal (Alt+F2 e digitar gnome-terminal)

11. Atualizar o Ubuntu

- sudo apt update
- sudo apt upgrade

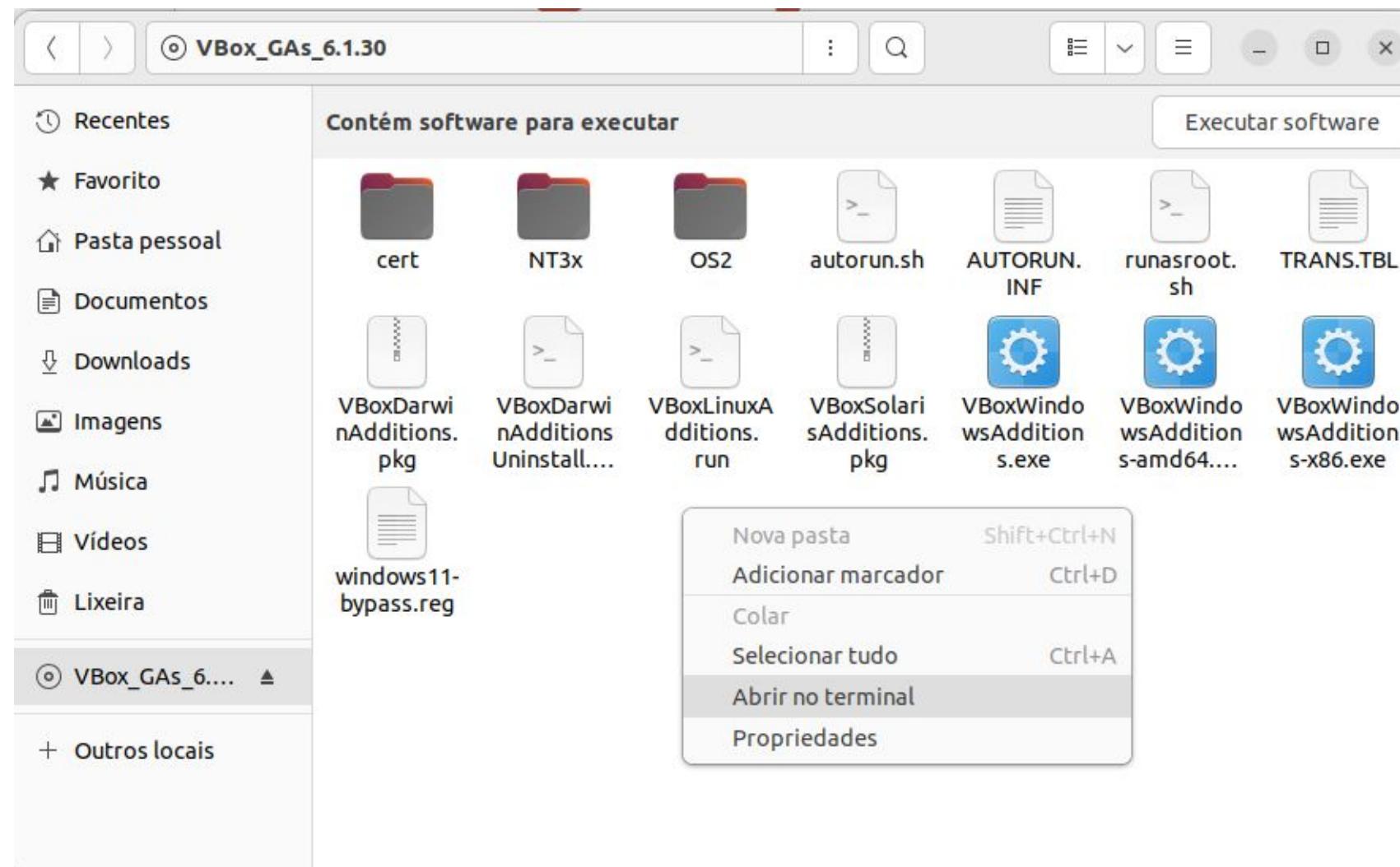


# Configuração da VM Ubuntu

## 12. Instalar add-ons (para melhor experiência com o Ubuntu)

```
➤ sudo apt install -y build-essential  
linux-headers-$ (uname -r)
```

## 13. Abrir no terminal a pasta com a mídia VBox\_GAs



## 14. Rodar o script para instalação

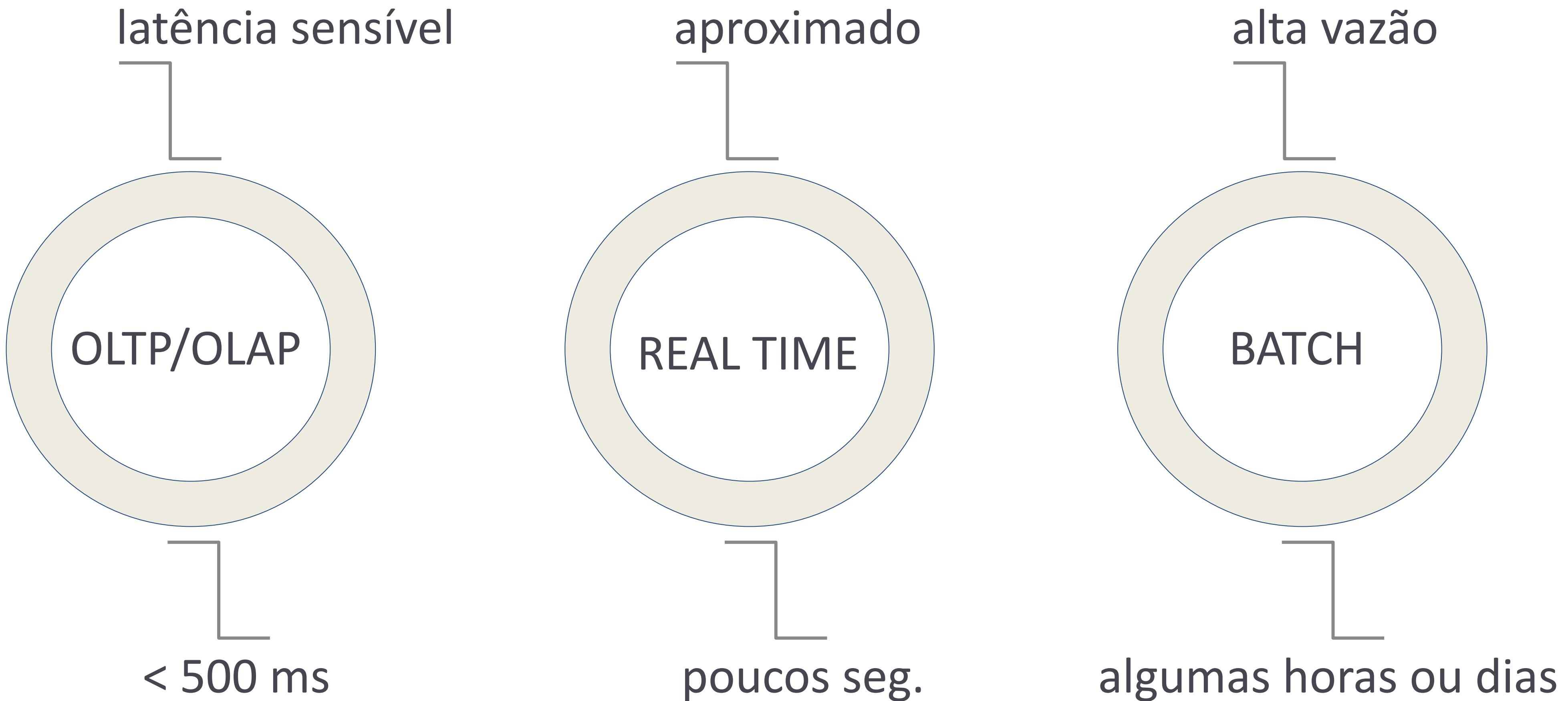
```
➤ ./autorun.sh
```

## 15. Reiniciar o Ubuntu

# O que é tempo real?

Milissegundos, segundos, minutos?

# O que é Tempo Real?



# O que é Tempo Real?

## REAL TIME TRENDS



Emerging break out  
trends in Twitter (in the  
form #hashtags)

## REAL TIME CONVERSATIONS



Real time sports  
conversations related  
with a topic (recent goal  
or touchdown)

## REAL TIME RECOMMENDATIONS



Real time product  
recommendations based  
on your behavior &  
profile

## REAL TIME SEARCH



Real time search of  
tweets with a budget <  
200 ms

Fonte: Real-Time Analytics with Apache Storm - <https://www.udacity.com/course/ud381>

# Problemas em streaming

1. Como obter dados a partir de várias fontes em tempo real?
2. Como processar esses dados?



Finalizando  
nossa Setup...

# Instalando algumas libs

## Instalar o curl

```
➤ sudo apt install curl
```

## Instalar o VSCode

```
➤ sudo snap install --classic code
```

## Instalar o Python

```
➤ sudo apt install python3-pip
```

## Instalar o Java

```
➤ sudo apt install default-jdk
```

Deixar instalando...

# Apache Kafka

# Apache Kafka

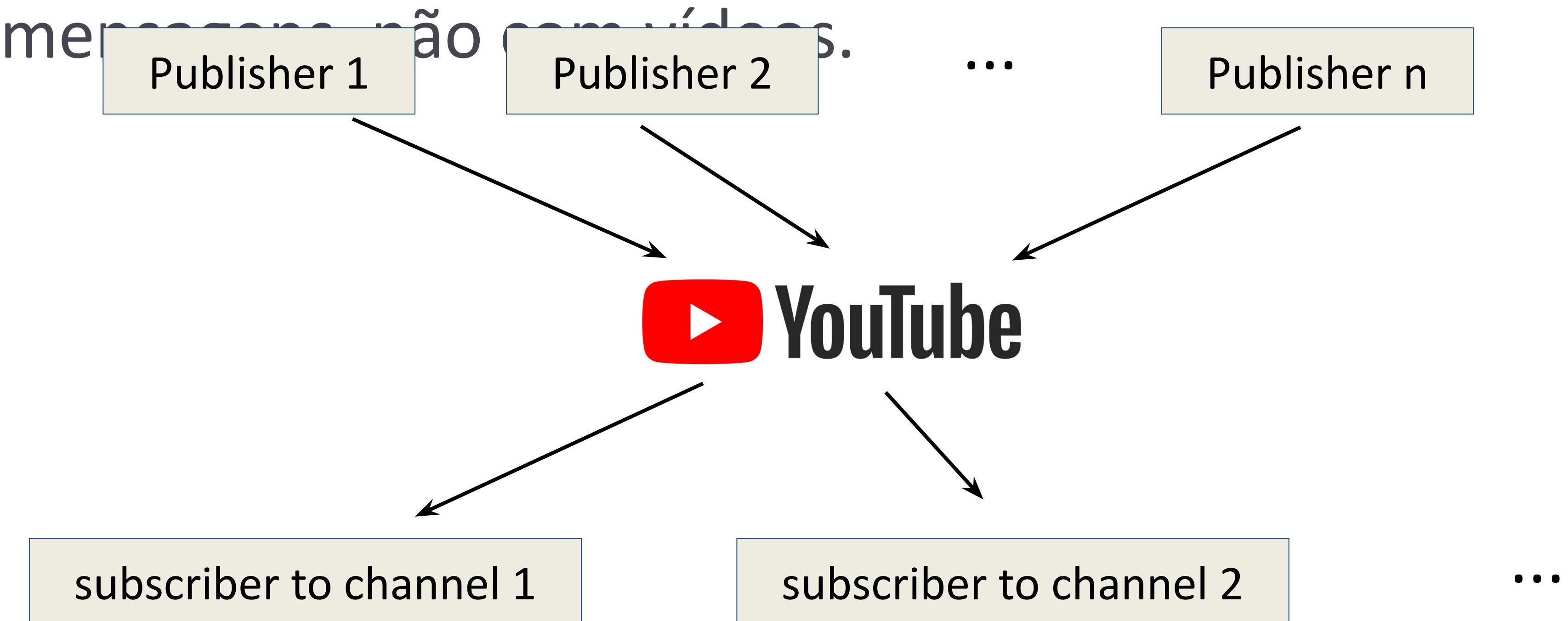
- Sistema de mensagens
  - Distribuído
  - Com alta vazão (*throughput*)
  - De geração (publicação) e leitura (sub-inscrição)
- Principais casos de uso:
  - Agregação de log
  - Mover/transformar conjuntos de dados em tempo real
  - Monitoramento

# Apache Kafka

- Originalmente desenvolvido pelo LinkedIn.
- Implementado em scala/Java.
- *Producers & Consumers.*
- Mensagens são associadas a tópicos, os quais representam um stream específico.
  - Logs web
  - Dados de sensores
- *Consumers* se inscrevem em um ou mais tópicos.

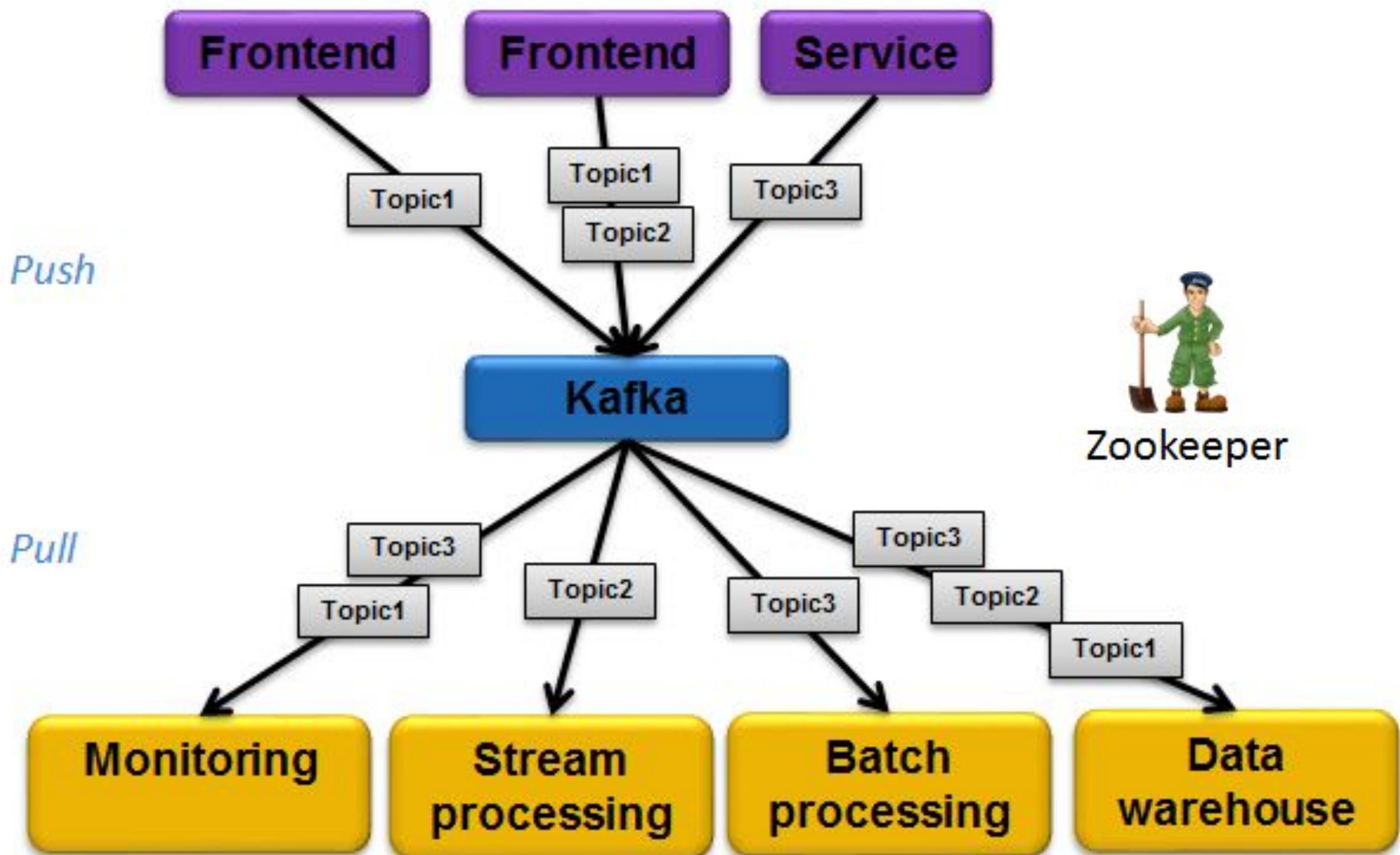
# Publisher-subscriber system

- Kafka pode ser visto como um sistema publisher/subscriber, como o Youtube, mas com mensagens não como vídeos.



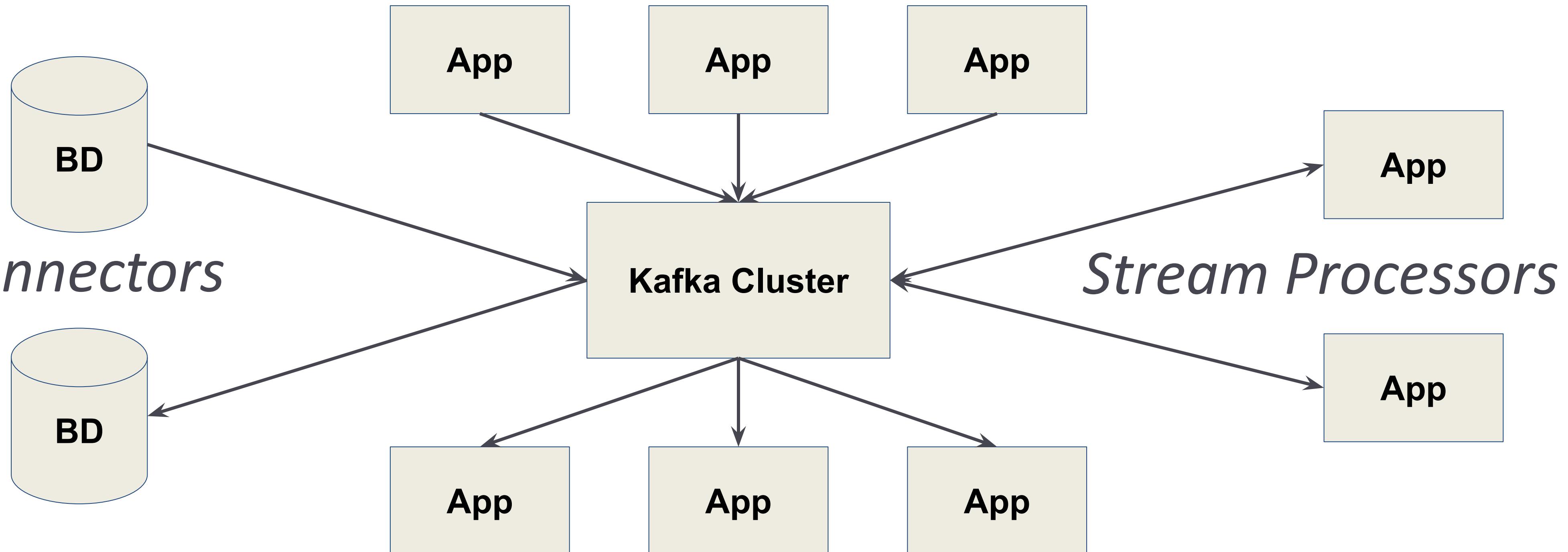
# Kafka: conceitos

Producers



# Kafka: arquitetura

*Producers*



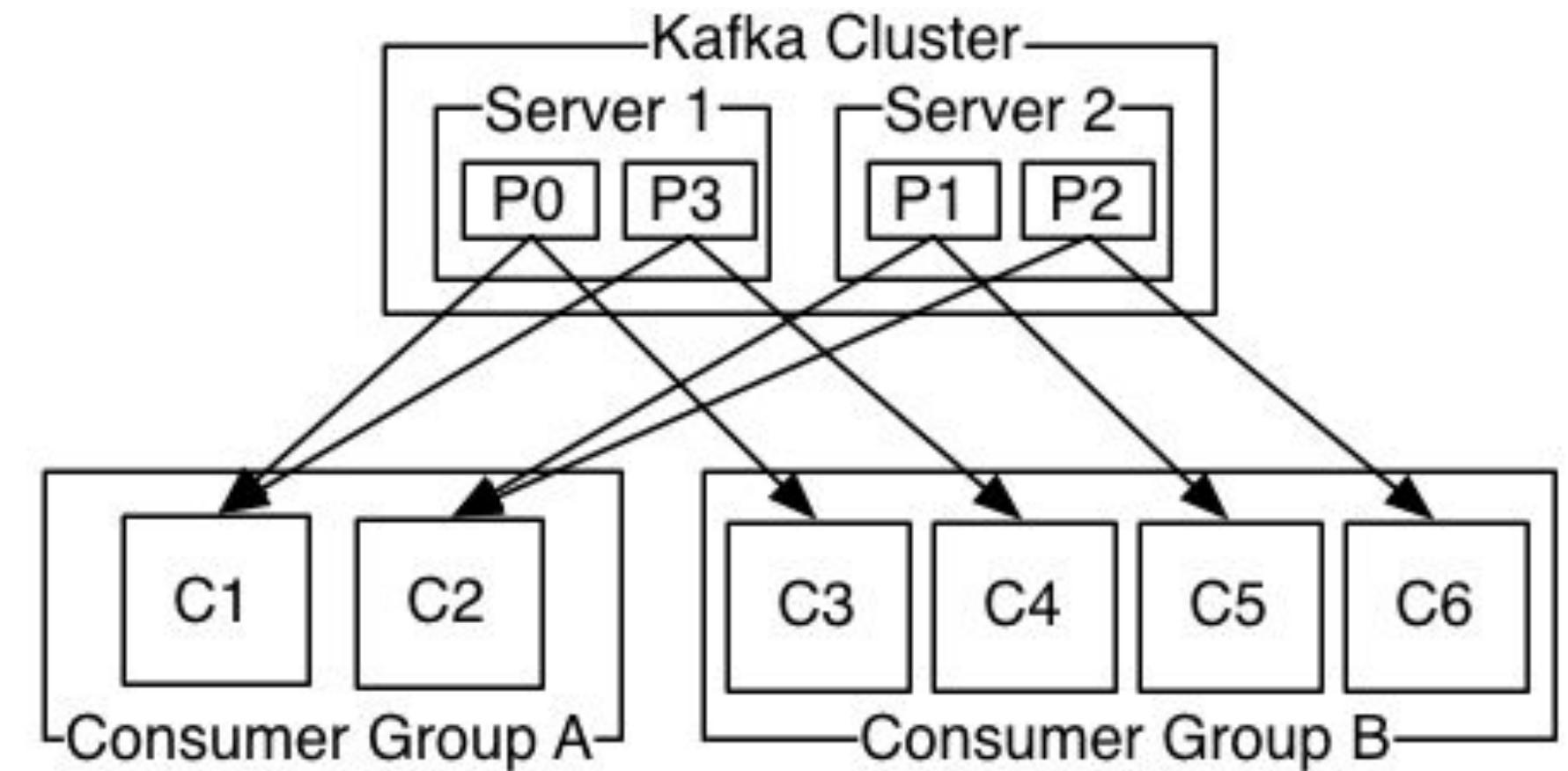
*Connectors*

*Stream Processors*

*Consumers*

# Kafka: escalabilidade

- Kafka pode ser distribuído entre muitos processos em vários servidores.
- *Consumers* também podem ser distribuídos.
- Tolerante a falhas.



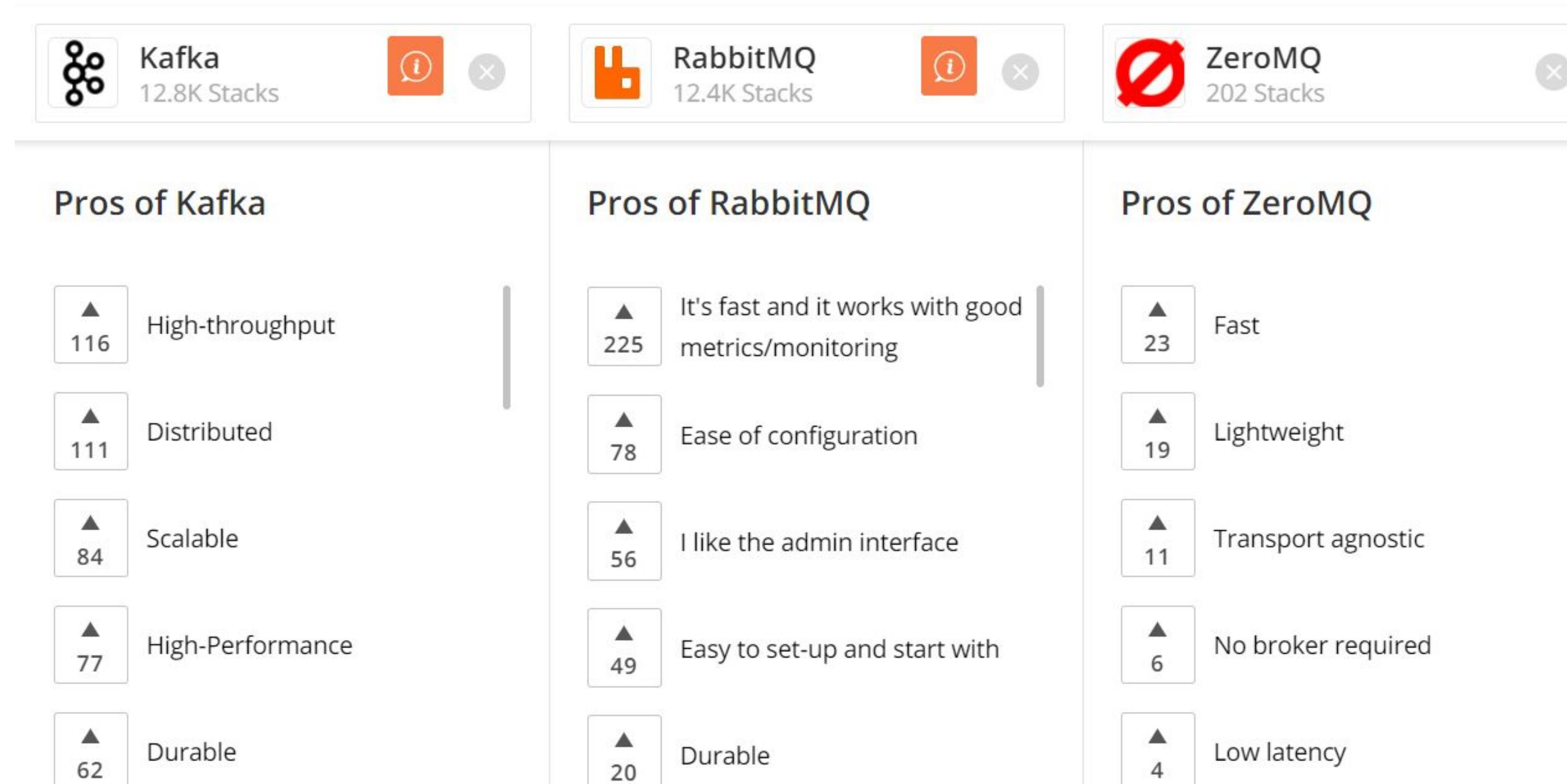
Fonte: <https://kafka.apache.org/intro.html>

# Kafka: pontos a considerar

- Simples sistema de mensagens, não de processamento.
- Não vive sem o **Zookeeper**, o qual pode se tornar um gargalo quando o número de tópicos/partições é muito grande ( $>>10000$ ).

# Kafka: quando comparado a outros concorrentes

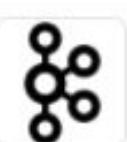
## Prós



Fonte: <https://stackshare.io/stackups/kafka-vs-rabbitmq-vs-zeromq>

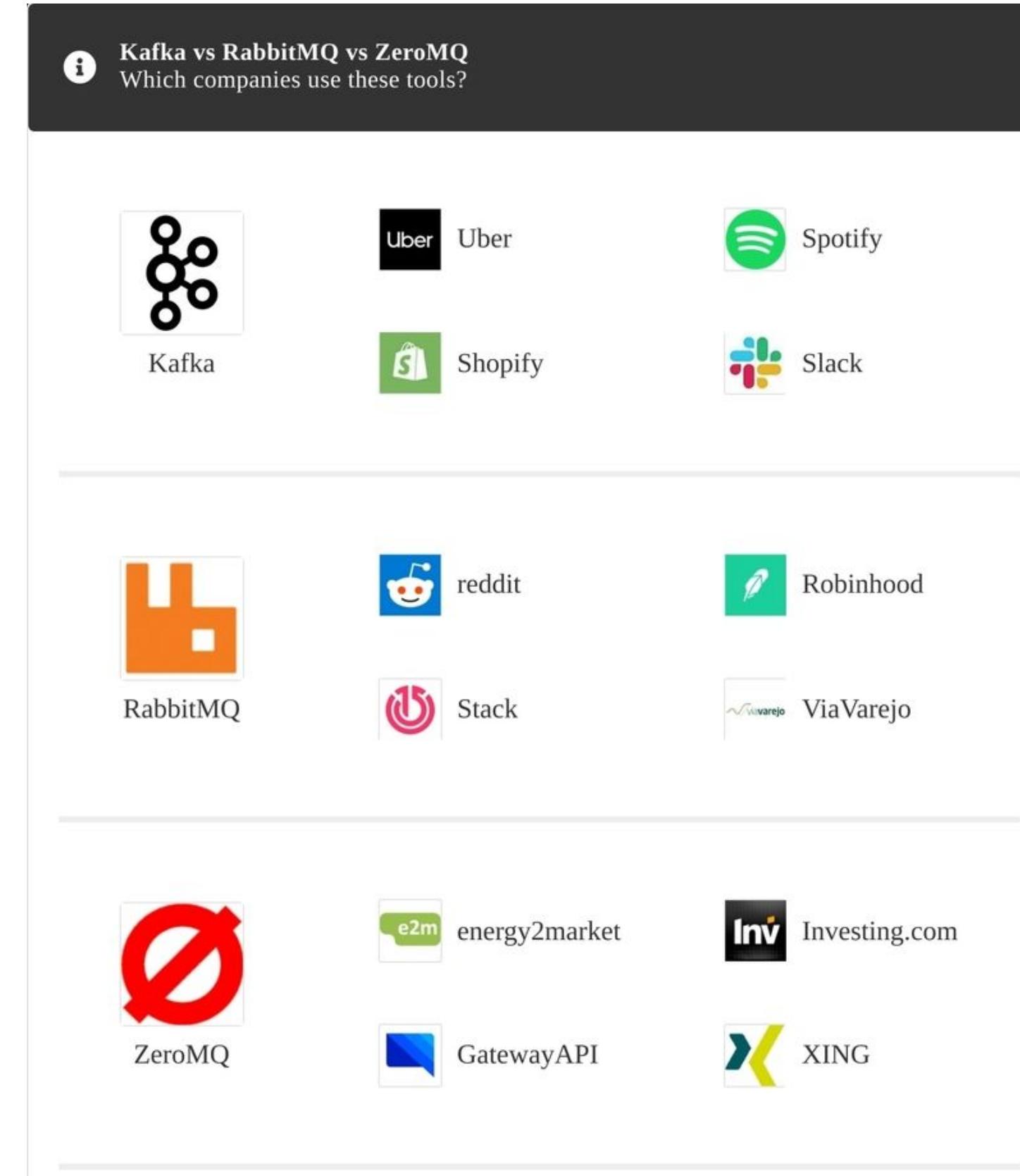
# Kafka: quando comparado a outros concorrentes

## Contras

 Kafka 12.8K Stacks	 RabbitMQ 12.4K Stacks	 ZeroMQ 202 Stacks
<h3>Cons of Kafka</h3> <ul style="list-style-type: none"><li>▲ 26 Non-Java clients are second-class citizens</li><li>▲ 25 Needs Zookeeper</li><li>▲ 7 Operational difficulties</li><li>▲ 1 Terrible Packaging</li></ul>	<h3>Cons of RabbitMQ</h3> <ul style="list-style-type: none"><li>▲ 9 Too complicated cluster/HA config and management</li><li>▲ 6 Needs Erlang runtime. Need ops good with Erlang runtime</li><li>▲ 5 Configuration must be done first, not by your code</li><li>▲ 4 Slow</li></ul>	<h3>Cons of ZeroMQ</h3> <ul style="list-style-type: none"><li>▲ 5 No message durability</li><li>▲ 3 Not a very reliable system - message delivery wise</li><li>▲ 1 M x N problem with M producers and N consumers</li></ul>

Fonte: <https://stackshare.io/stackups/kafka-vs-rabbitmq-vs-zeromq>

# Kafka: quando comparado a outros concorrentes



Fonte: <https://stackshare.io/stackups/kafka-vs-rabbitmq-vs-zeromq>

# Dúvidas?