

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 os dados são produzidos.

Resolver problemas do mundo real relacionados a **Streaming de Dados em tempo real**.

Ementa (dia 1)

- Introdução a Streaming de Dados
- Como **obter** os dados de Streaming
- Setup do ambiente

Ementa (dia 2)

- Apache Kafka
- Como processar os dados de Streaming
- Apache Spark
- Resolução de problemas com Spark

Ementa (dia 3)

- Resolução de problemas de mineração de dados de Streaming com Python e Spark:
 - Filtragem
 - Estatísticas
 - Janelas

Ementa (dia 4)

- Spark Streaming Project

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 (em dupla)

Recursos

Lista de e-mails:

uni7-ciencia-de-dados-turma6@googlegroups.com

m

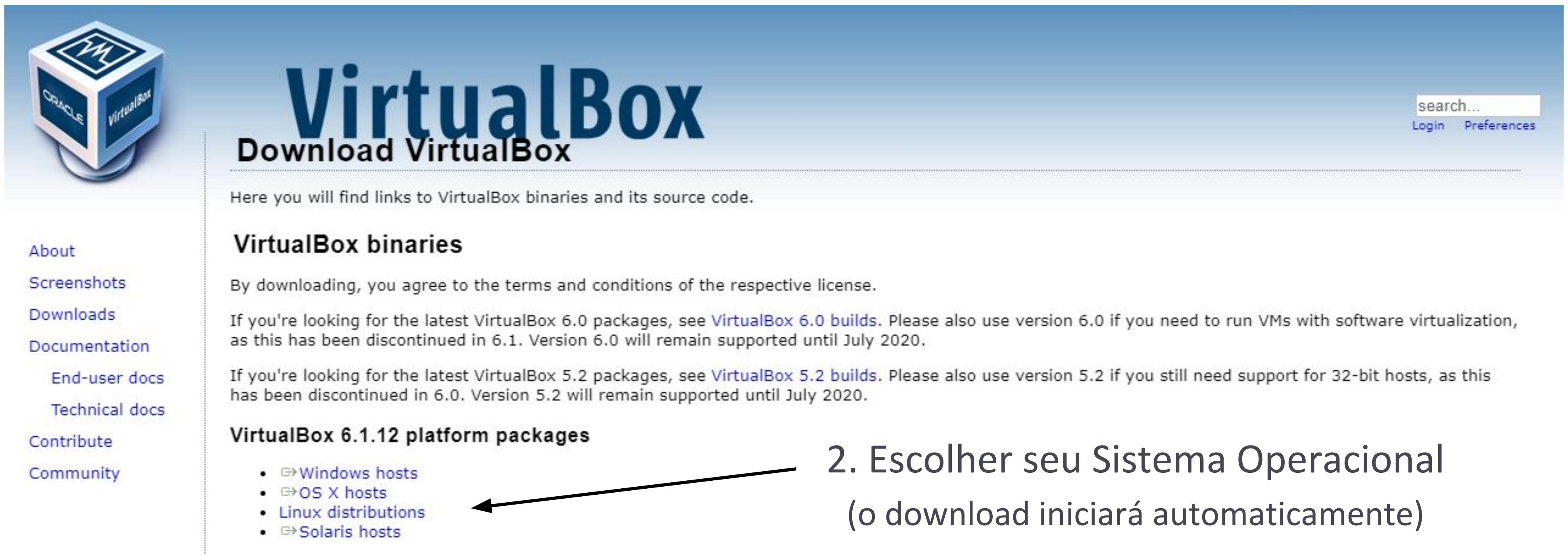
Arquivos no Microsoft Teams:

Slides, dados, scripts, livros e artigos

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 logo, which is a blue cube with 'ORACLE' on the front face and 'VirtualBox' on the right face. The main title 'VirtualBox' is in large blue letters, with 'Download VirtualBox' below it. To the right is a search bar labeled 'search...' and buttons for 'Login' and 'Preferences'. On the left, there's a sidebar with links: 'About', 'Screenshots', 'Downloads', 'Documentation', 'End-user docs', 'Technical docs', 'Contribute', and 'Community'. The main content area starts with a note: 'Here you will find links to VirtualBox binaries and its source code.' Below this is a section titled 'VirtualBox binaries' with a note about accepting terms and conditions. It mentions that version 6.0 is supported until July 2020. Another note says version 5.2 is supported until July 2020. A section titled 'VirtualBox 6.1.12 platform packages' lists four options: 'Windows hosts', 'OS X hosts', 'Linux distributions', and 'Solaris hosts'. A black arrow points from the text '2. Escolher seu Sistema Operacional' down to the 'Linux distributions' link.

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 6.1.12 platform packages

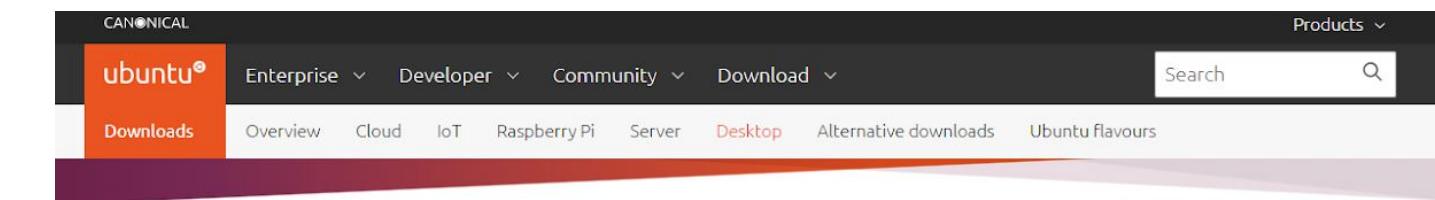
- [Windows hosts](#)
- [OS X hosts](#)
- [Linux distributions](#)
- [Solaris hosts](#)

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

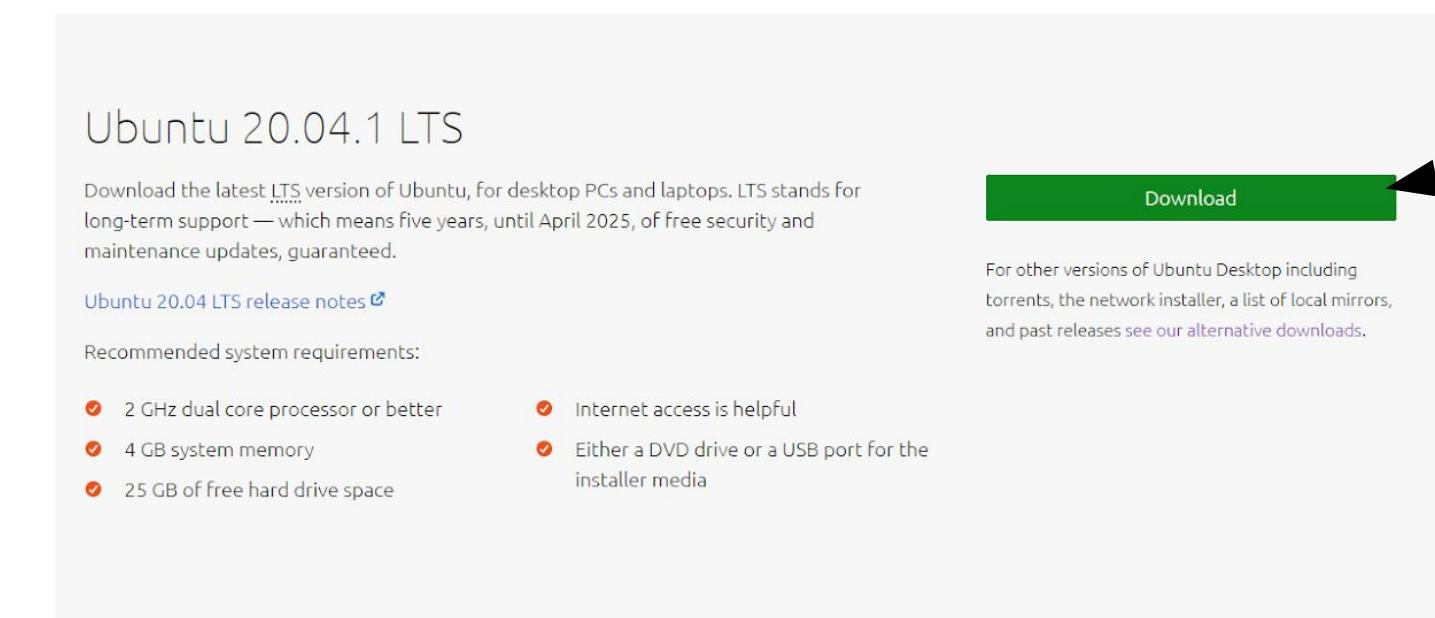
Download do Ubuntu

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

4. Realizar o download
do Ubuntu



Download Ubuntu Desktop



More ways to use Ubuntu everywhere



On-demand VMs for cloud devs on Windows, Mac and Linux desktops with Multipass.

[Get Multipass](#)



Install a Ubuntu terminal environment with Windows Subsystem for Linux (WSL).

[Get started with WSL](#)

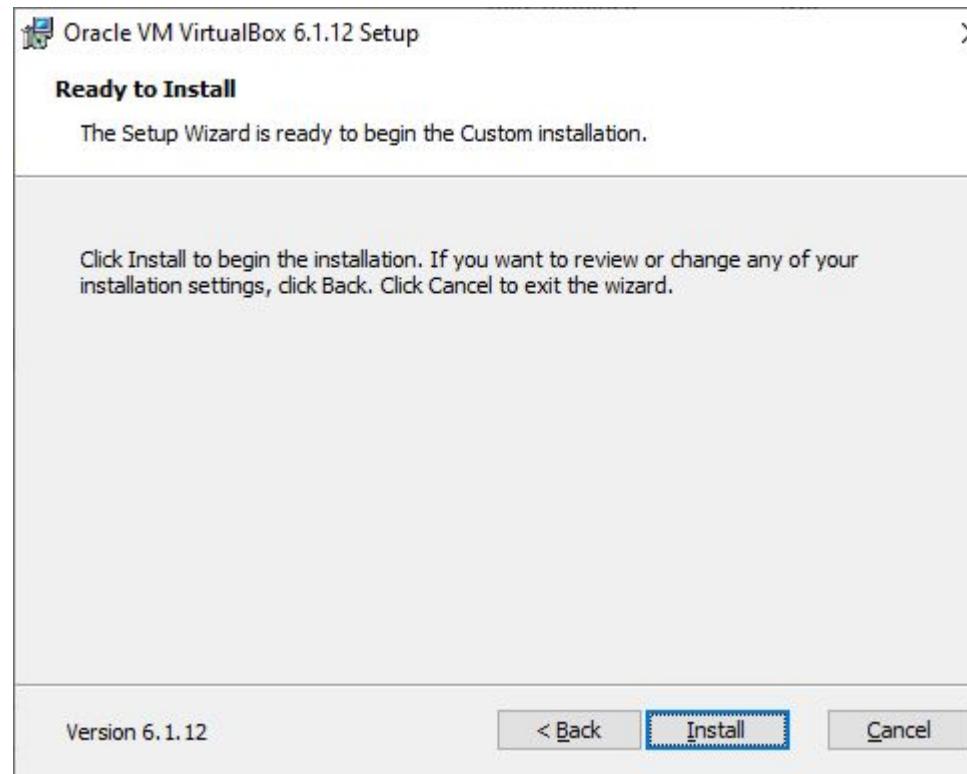
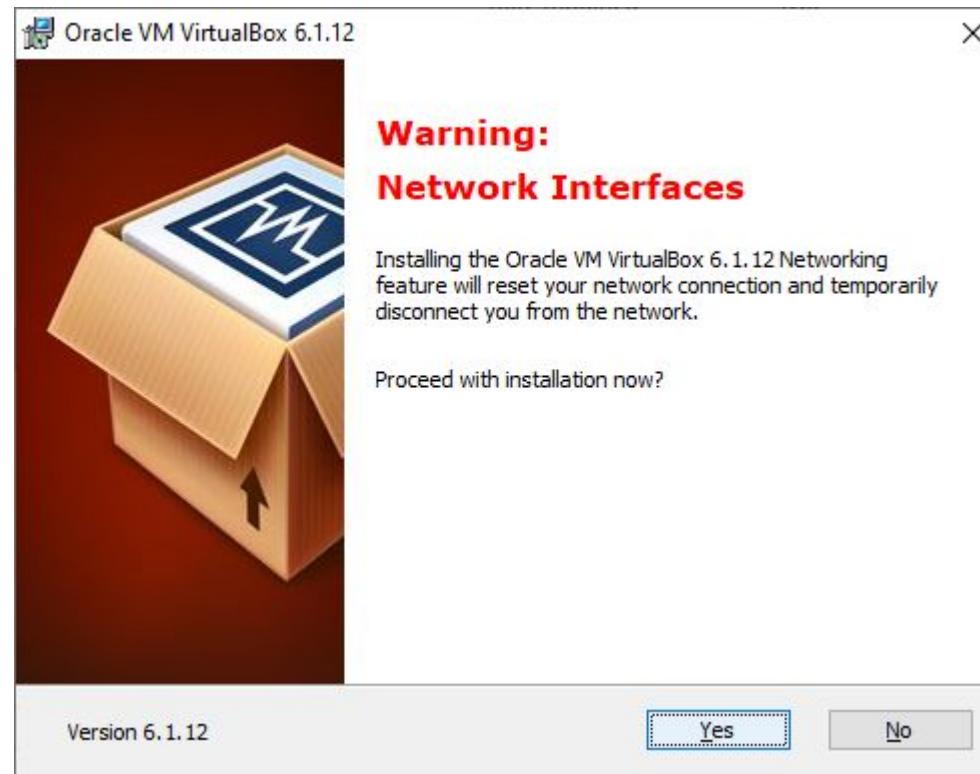
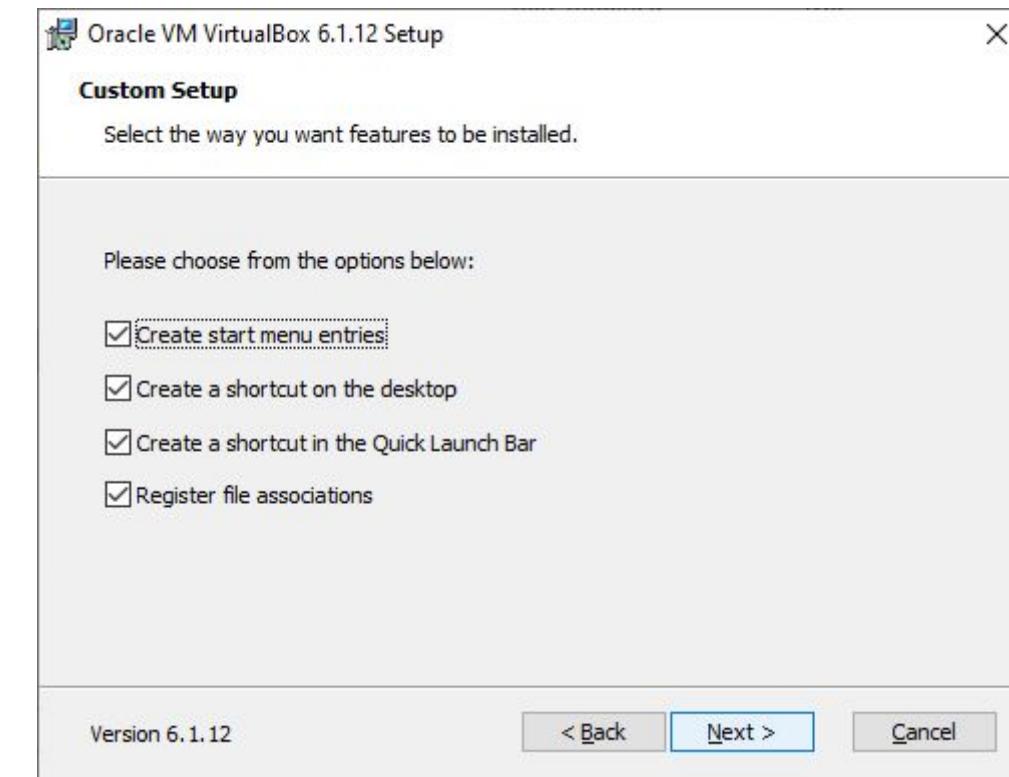
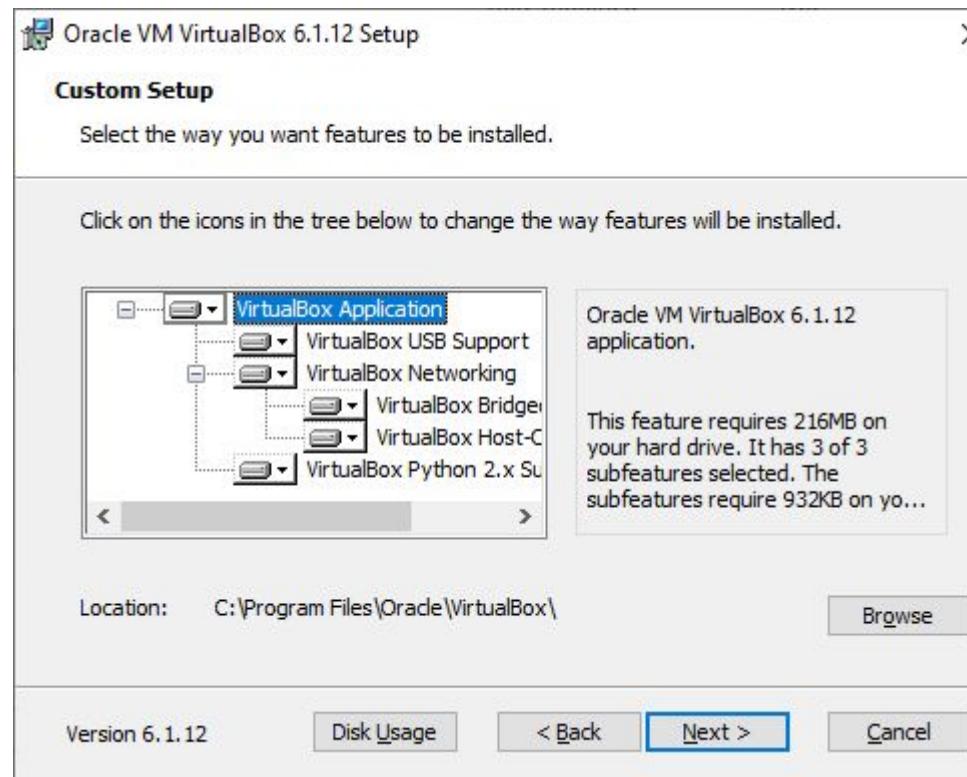


Download a free appliance image for your PC, Raspberry Pi or a virtual machine.

[Try an instant appliance now](#)

Instalação do VirtualBox

5. Instalar o VirtualBox



Criação de uma VM (Virtual Machine)

6. Criar uma nova Máquina Virtual Ubuntu

The screenshot shows the Oracle VM VirtualBox Manager interface. On the left, the main window displays a welcome message and global tools. In the center, four sub-windows guide the creation of a new virtual machine:

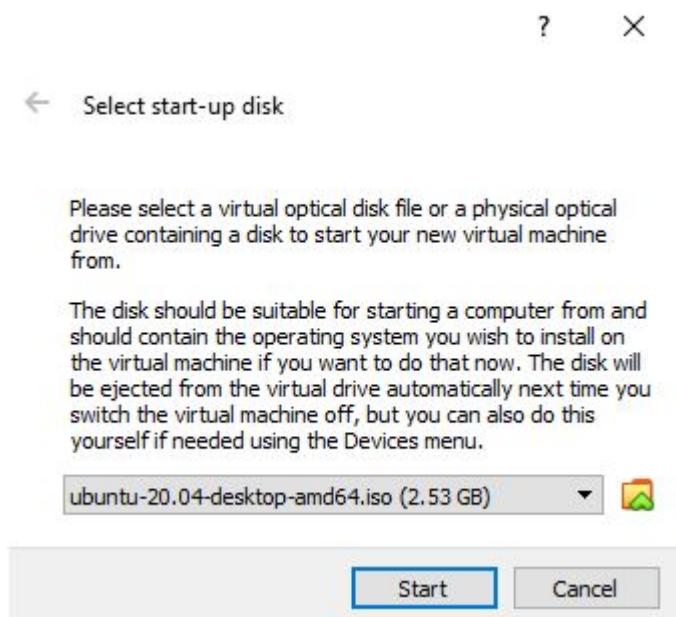
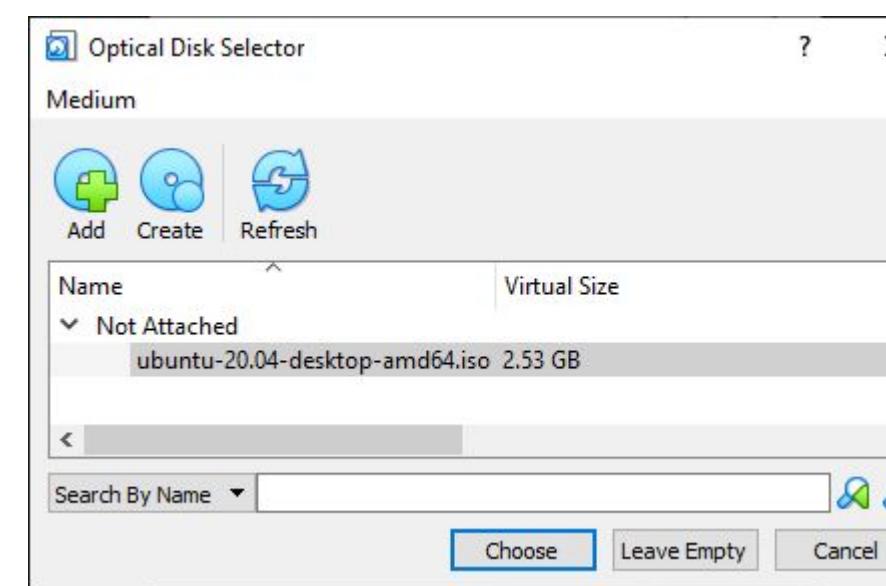
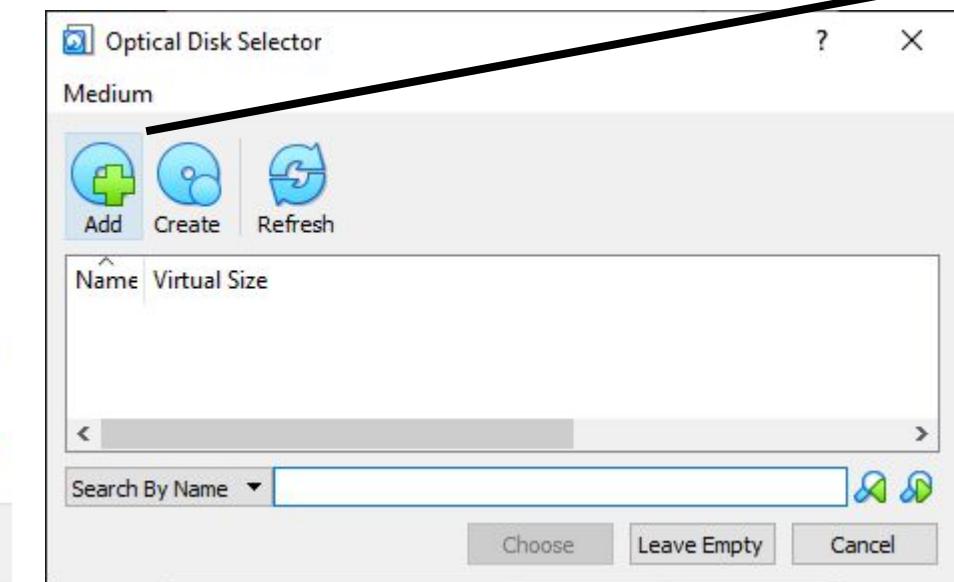
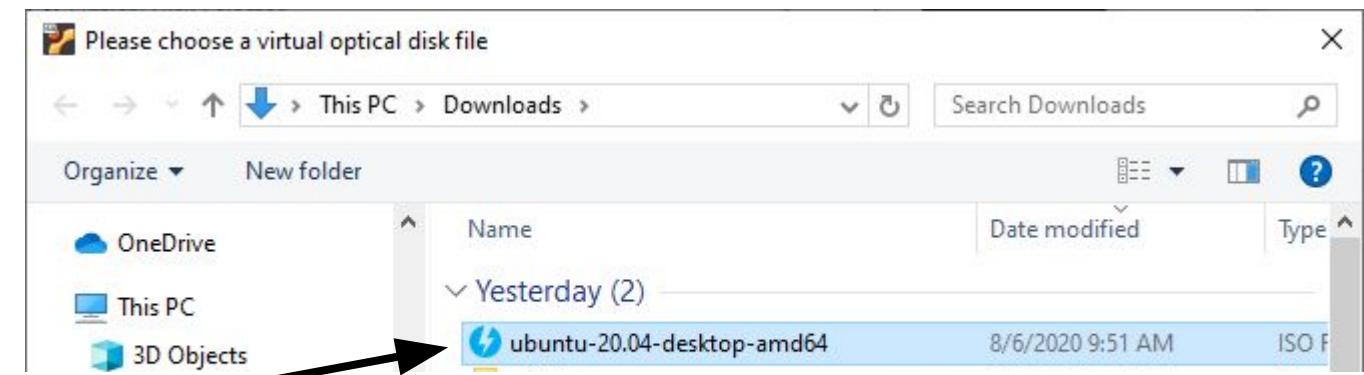
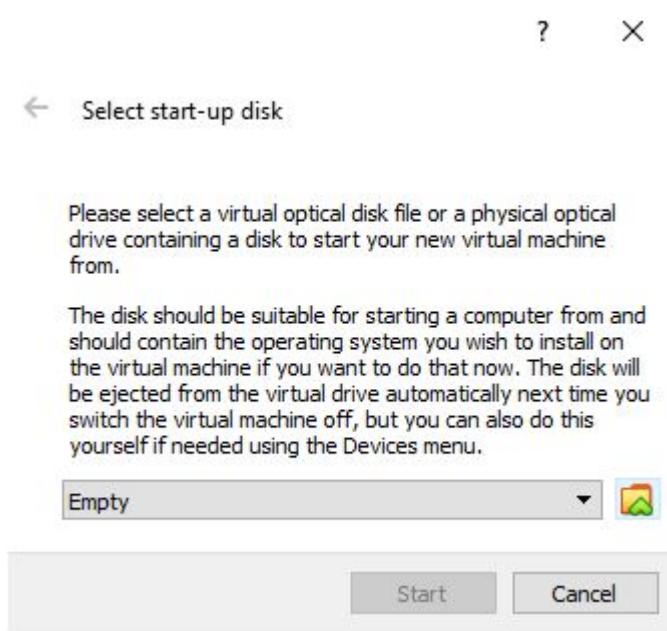
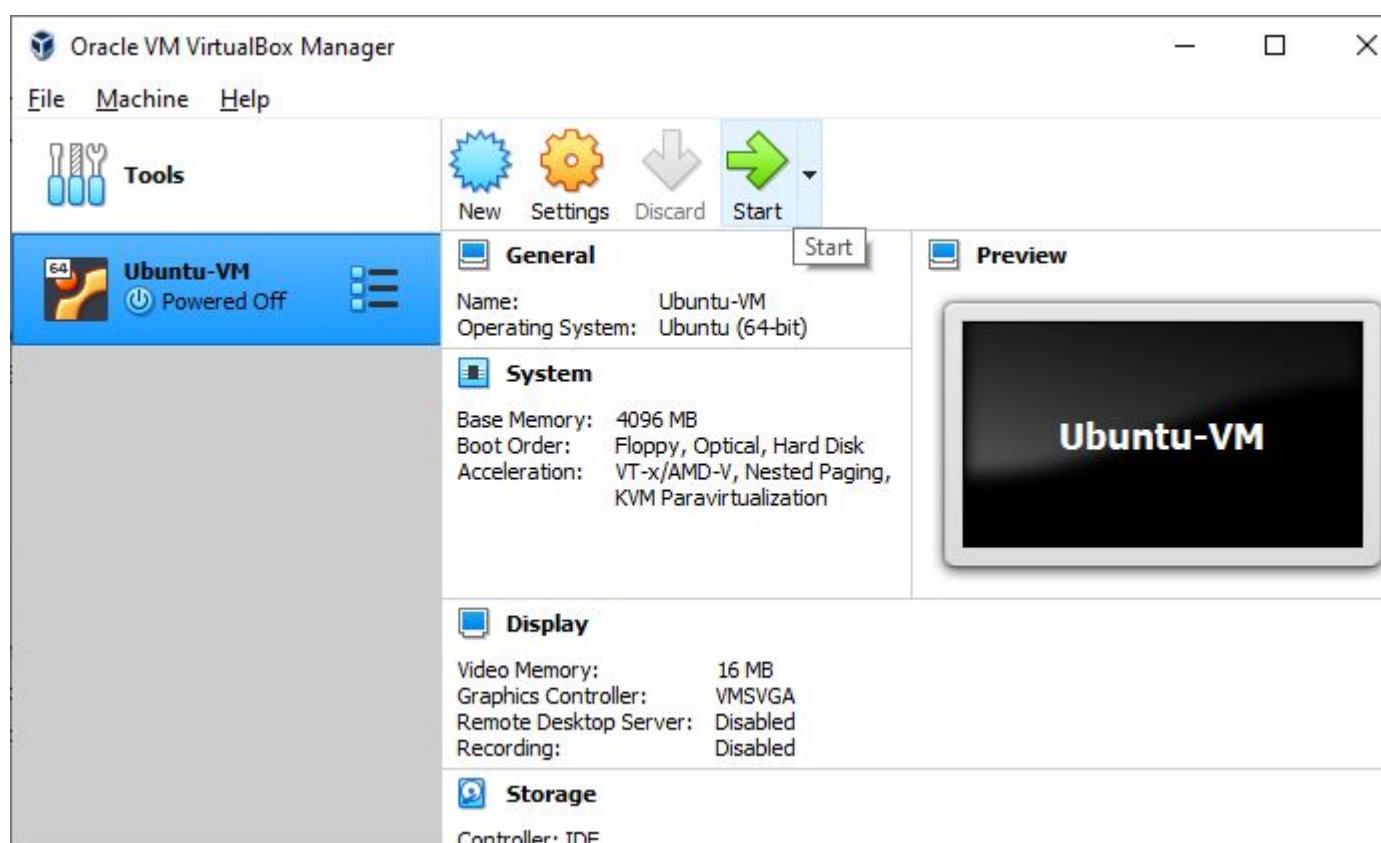
- Name and operating system**: Set Name to "Ubuntu-VM", Machine Folder to "VMs", Type to "Linux", and Version to "Ubuntu (64-bit)".
- Memory size**: Set memory to 1024 MB.
- Hard disk**: Set hard disk to "Create a virtual hard disk now".
- Create Virtual Hard Disk**: Set file type to "VDI (VirtualBox Disk Image)", storage to "dynamically allocated", and size to 25 GB.

On the right, the final configuration summary for "Ubuntu-VM" is shown, including settings for General, System, Display, and Storage.

General	System	Display	Storage
Name: Ubuntu-VM Operating System: Ubuntu (64-bit)	Base Memory: 4096 MB Boot Order: Floppy, Optical, Hard Disk Acceleration: VT-x/AMD-V, Nested Paging, KVM Paravirtualization	Video Memory: 16 MB Graphics Controller: VMSVGA Remote Desktop Server: Disabled Recording: Disabled	Controller: IDE

Criação de uma VM (Virtual Machine)

7. Iniciar a Máquina Virtual Ubuntu

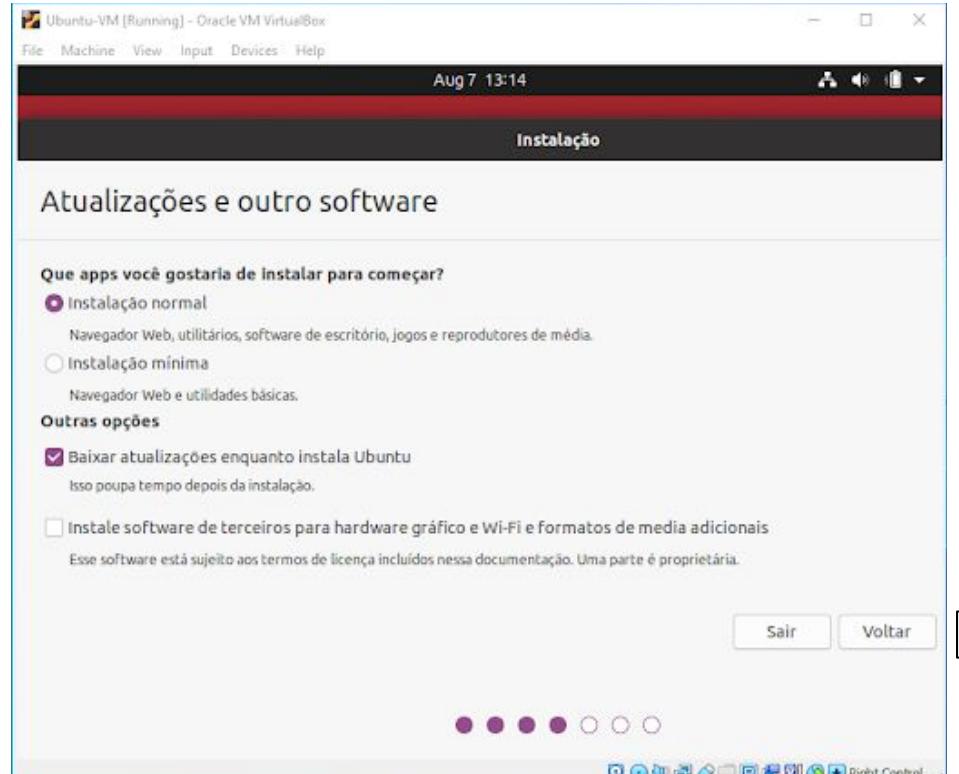
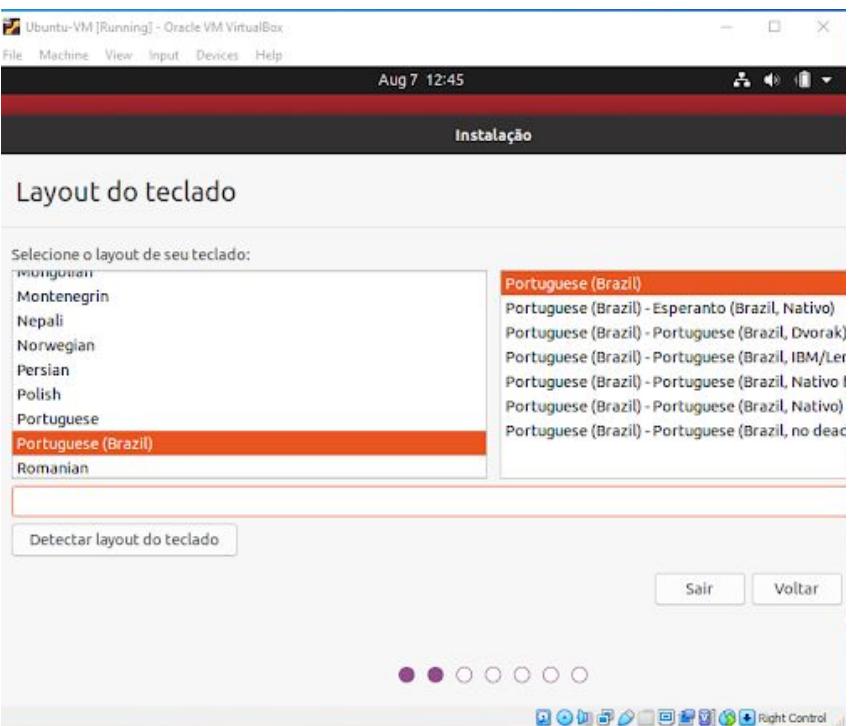


Instalação do Ubuntu

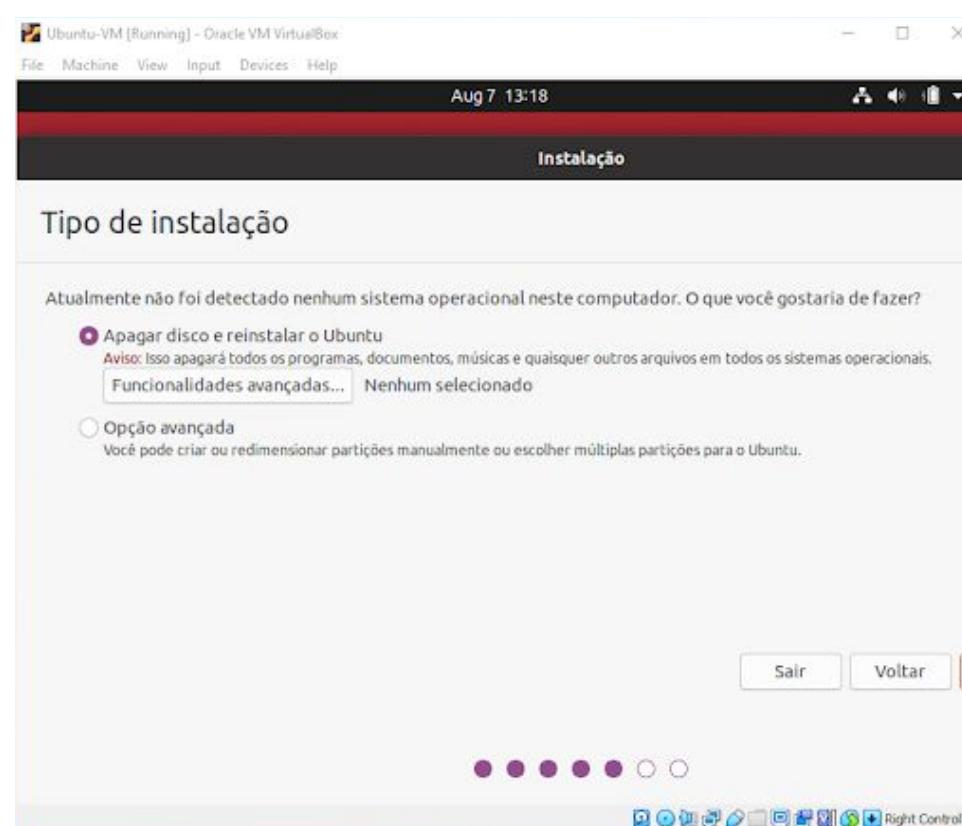
8. Instalar o Ubuntu na VM



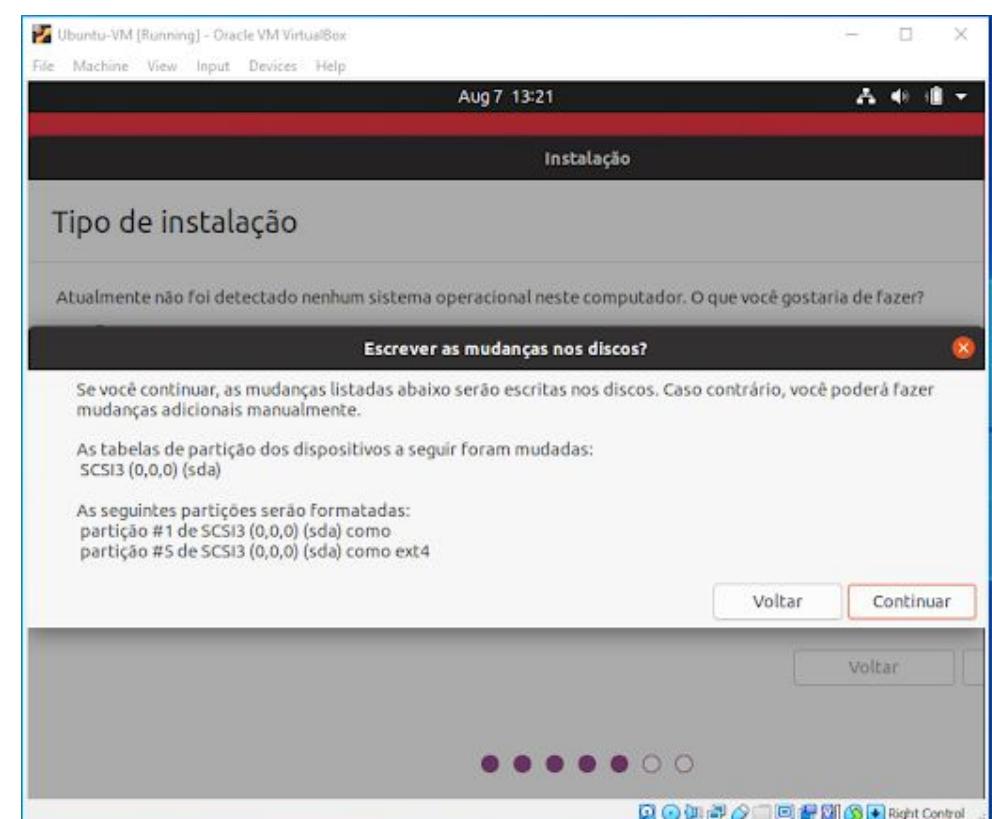
dois cliques



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

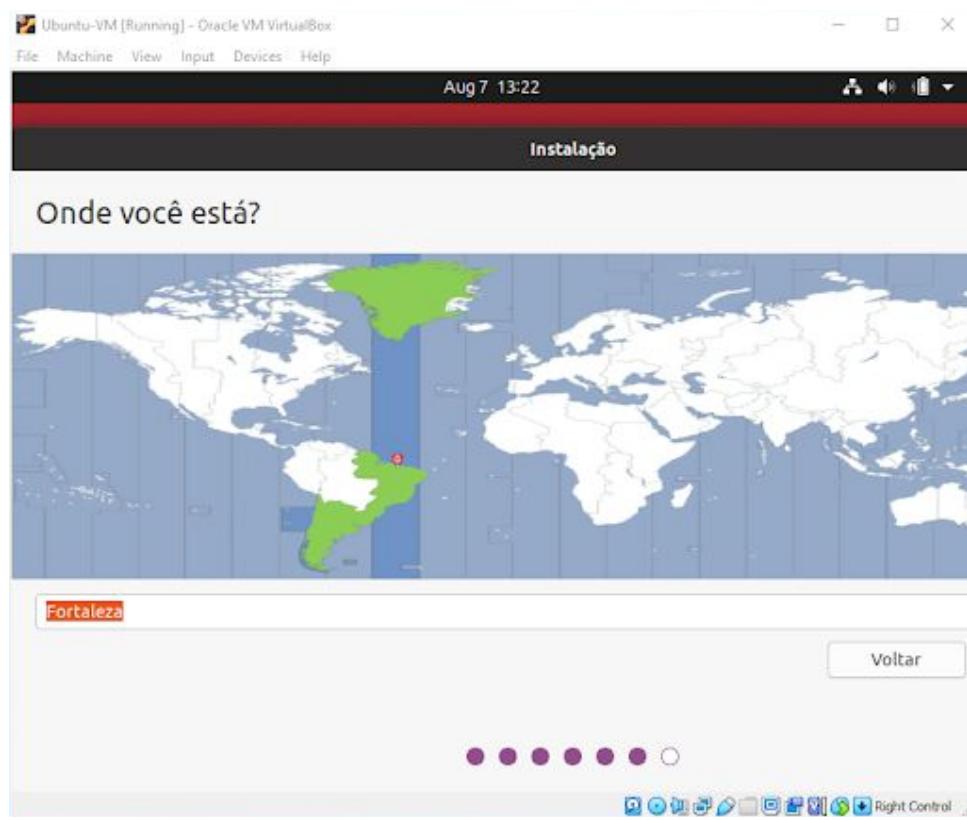


o botão 'continuar' pode
estar escondido aqui.
Acessá-lo via tecla 'tab'

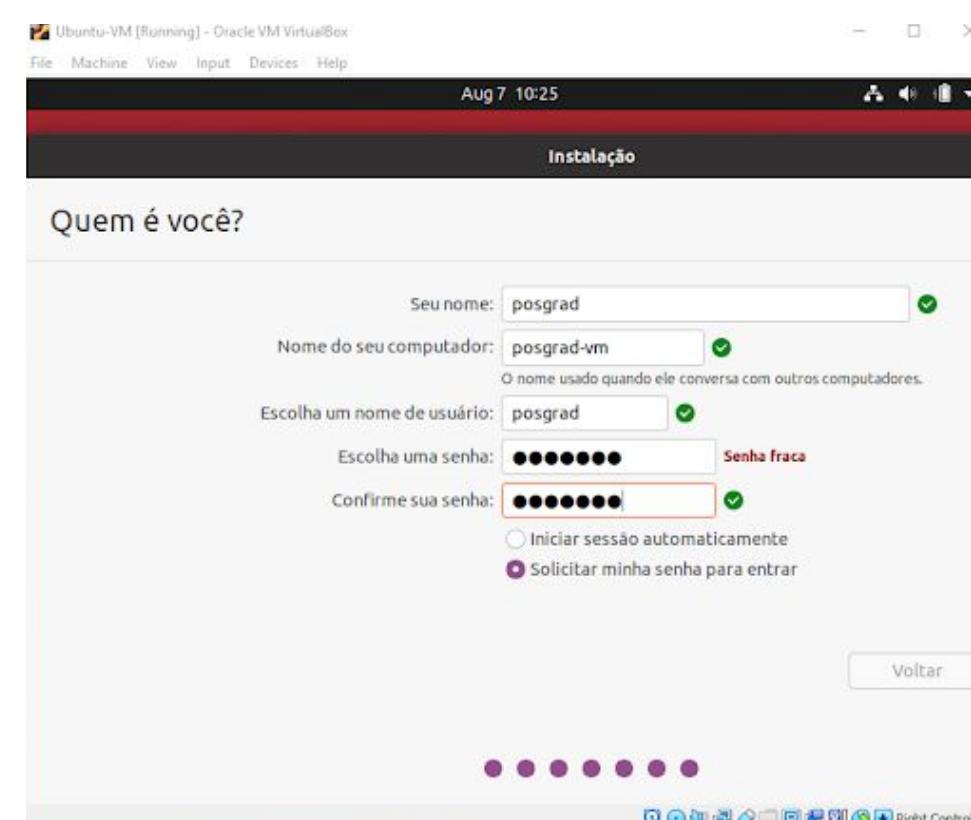


Instalação do Ubuntu

8. Instalar o Ubuntu na VM

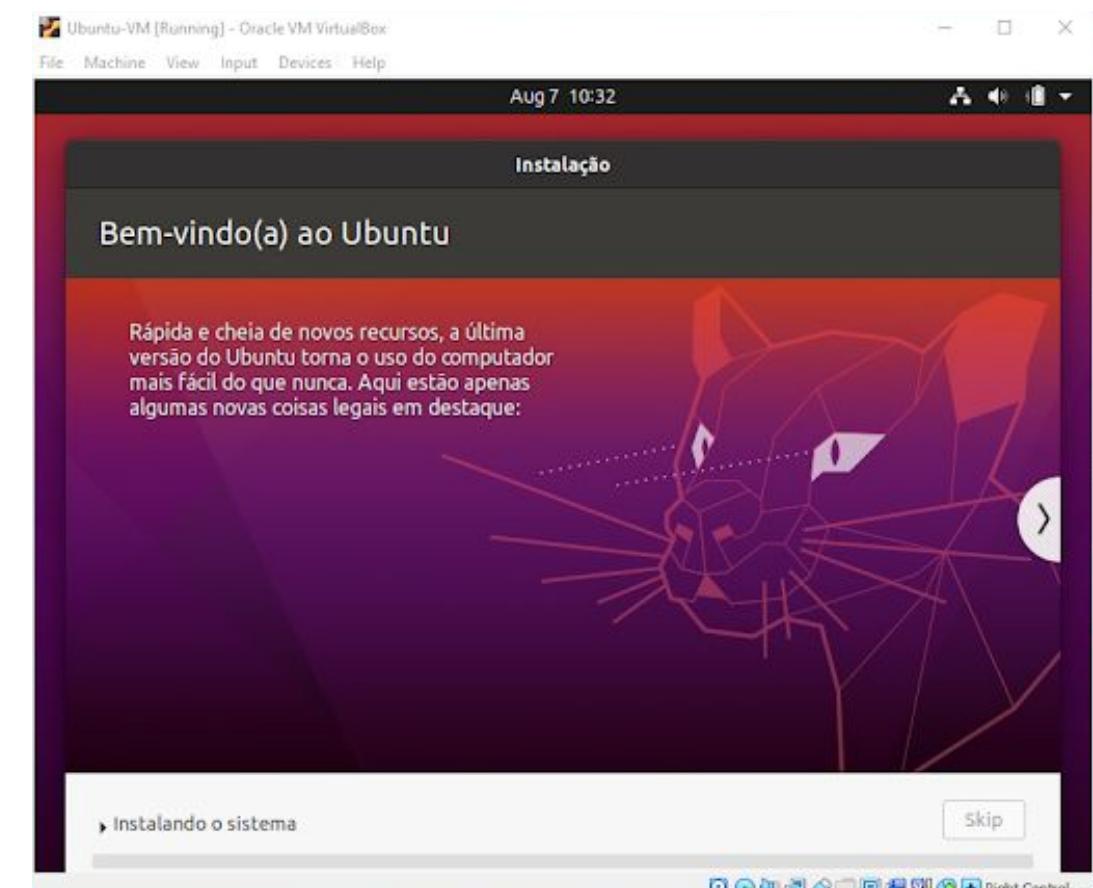


o botão 'continuar' pode
estar escondido aqui.
Acessá-lo via tecla 'tab'



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

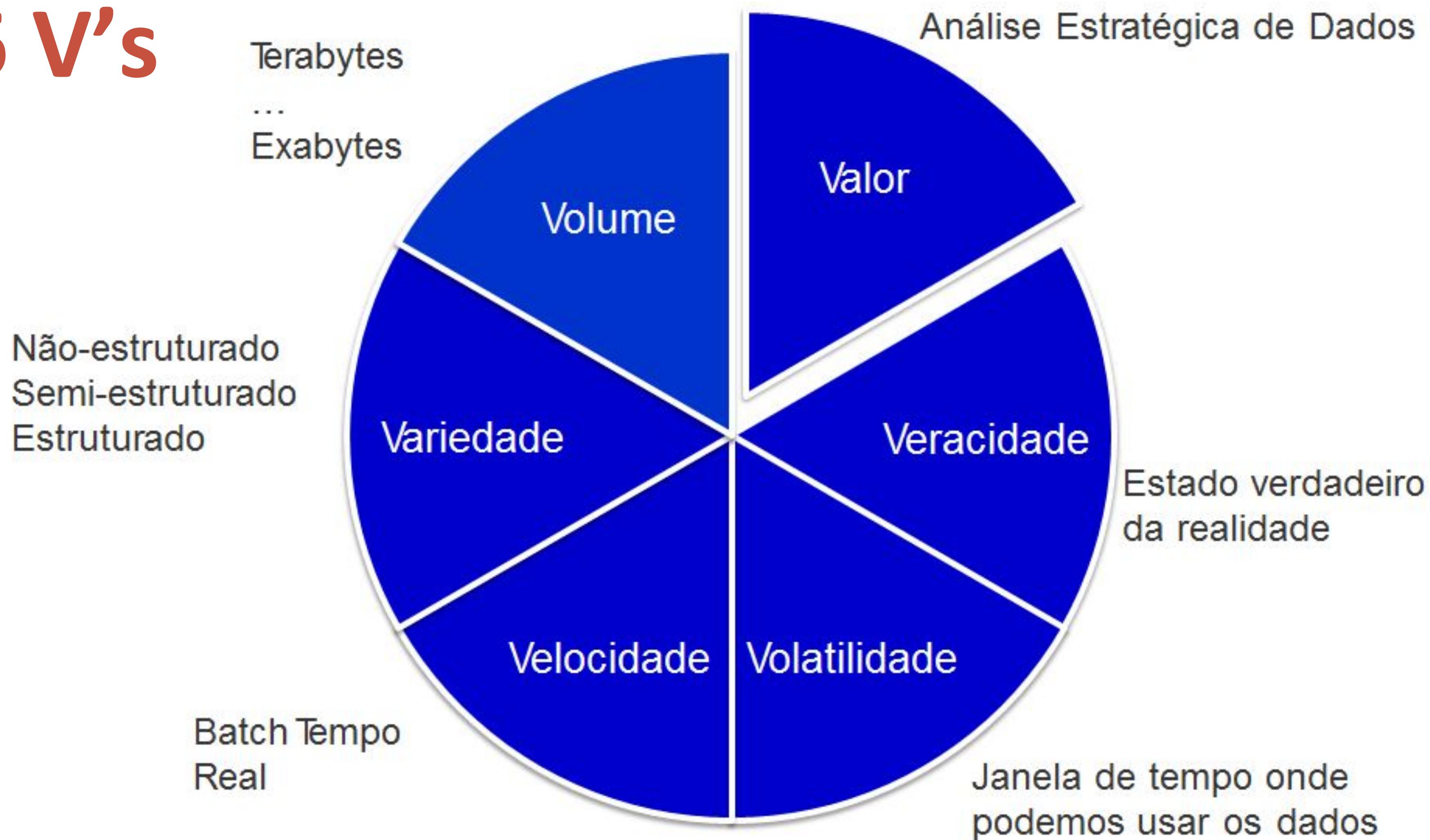
Big Data são dados que excedem o **armazenamento, o processamento e a capacidade dos sistemas convencionais**

Volume de dados muito grande

Dados são gerados rapidamente

Dados não se encaixam nas estruturas de arquiteturas de sistemas atuais

6 V's



Processamento em Batch



Todas as
entradas

Hadoop,
Spark, etc.

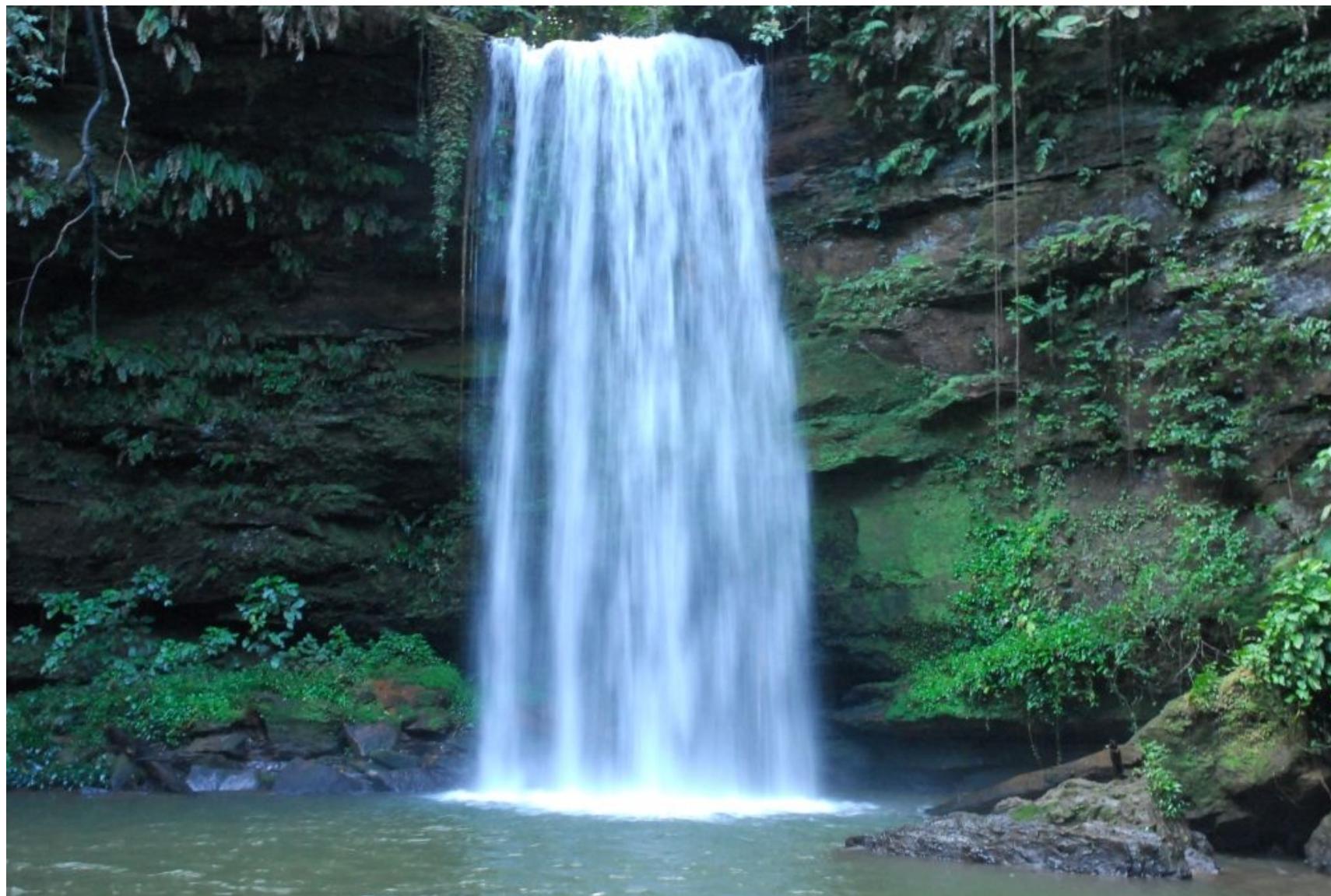
Todas as
saídas

Streaming

O que vem a sua cabeça quando
você escuta a palavra streaming?

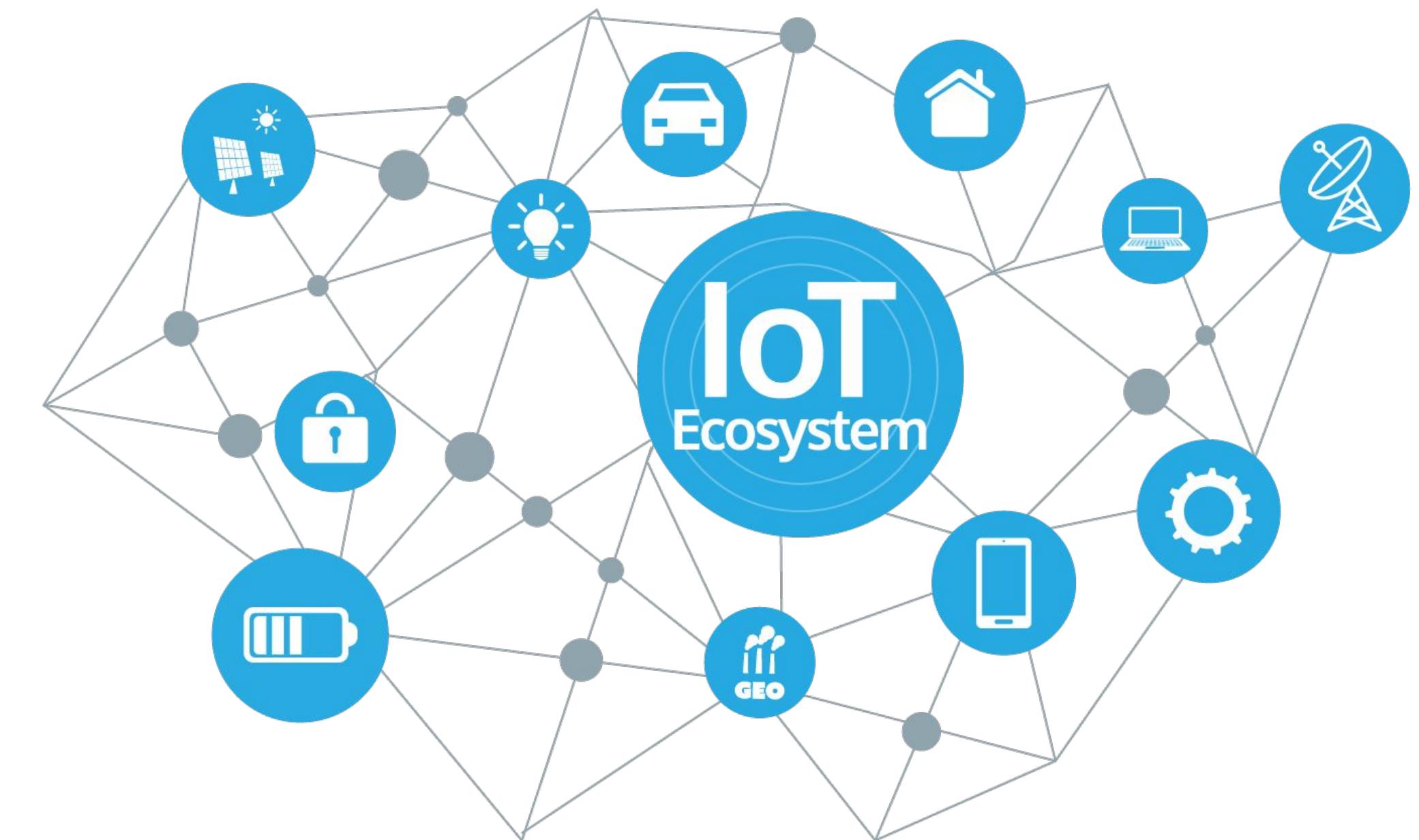
Streaming

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

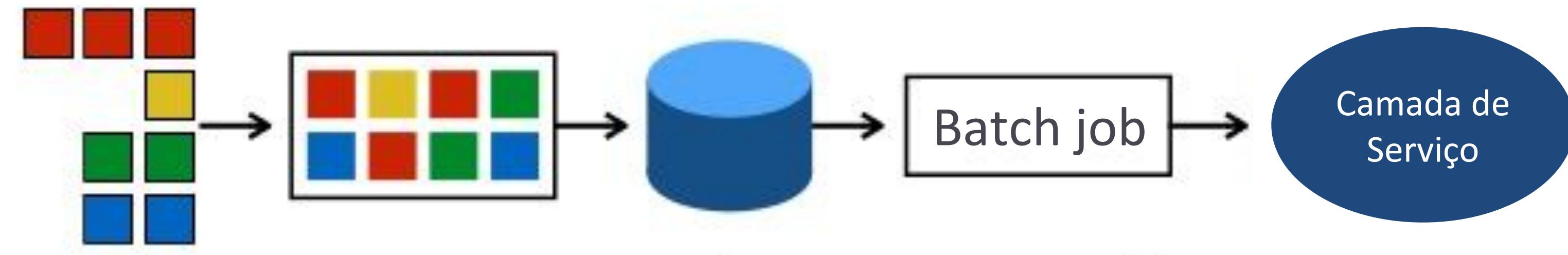
Dados em grafo: redes sociais

Dados de mobilidade: coordenadas geográficas x tempo

Etc.

Processamento de Streamings

Em geral:



continuamente
produzido

arquivos são
streams finitos

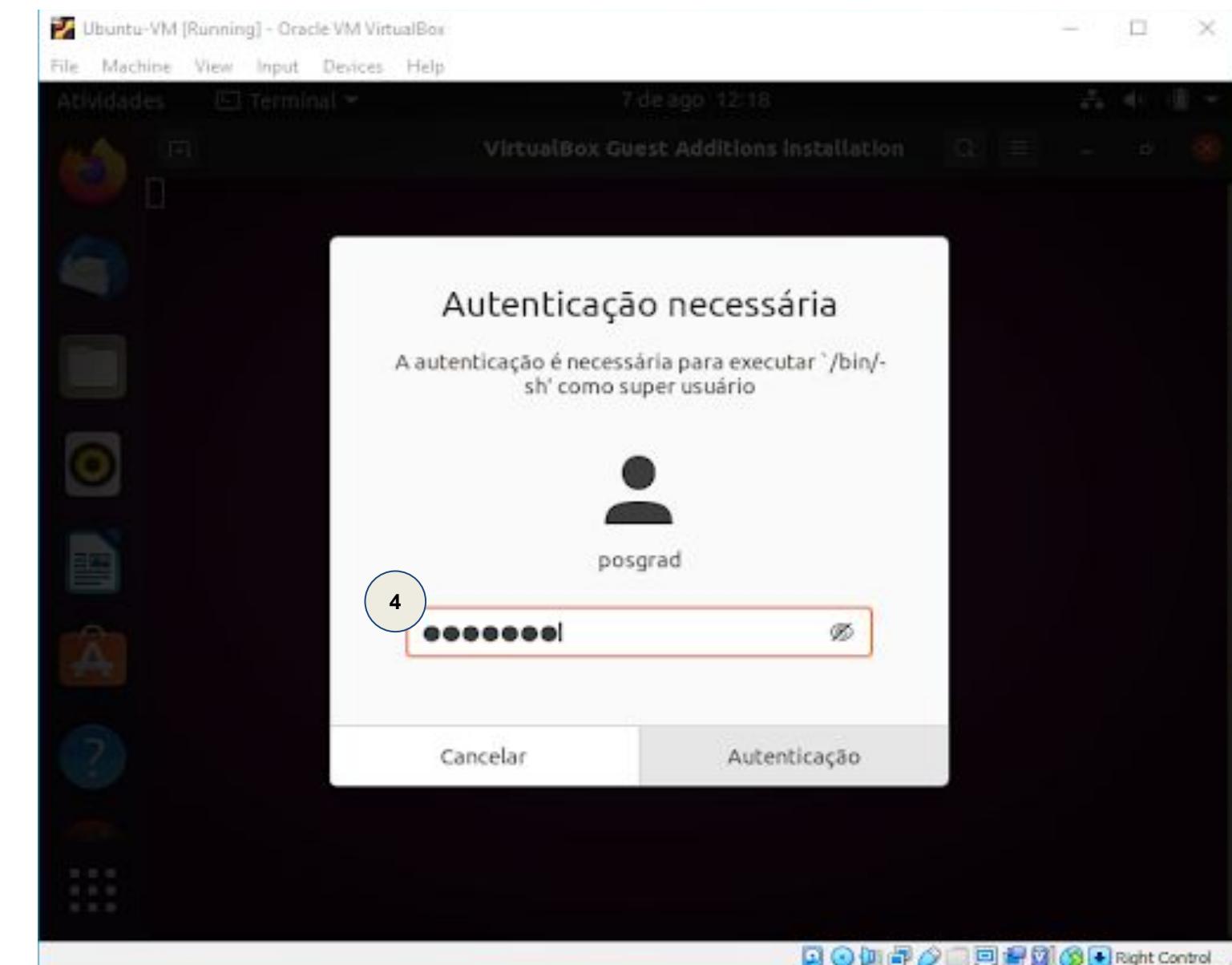
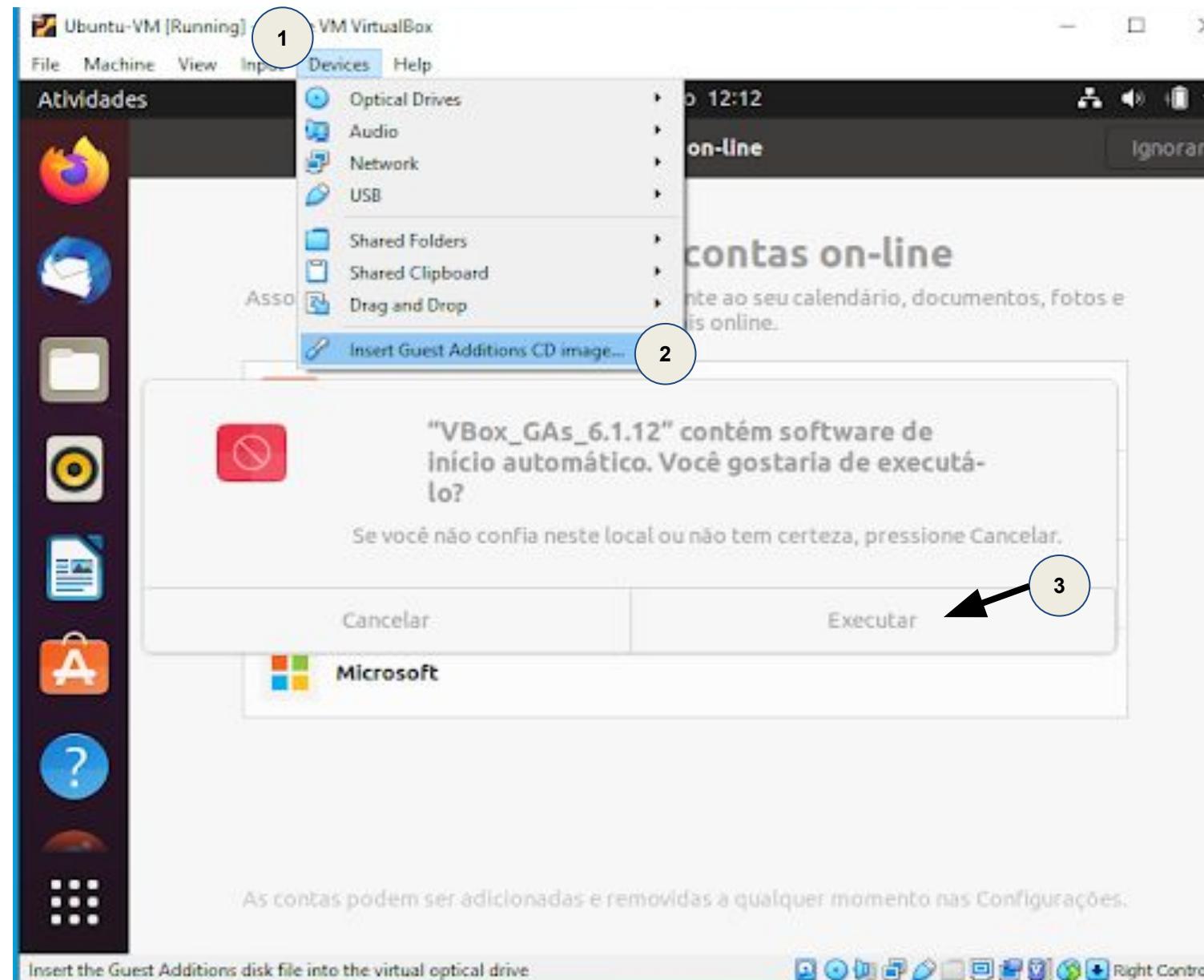
periodicamente
executado

Continuando
nossa Setup...

Configuração do Ubuntu

9. Logar na VM

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



11. Reiniciar a VM

Atualizando o Ubuntu

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

Atualizar o Ubuntu

- sudo apt update
- sudo apt upgrade

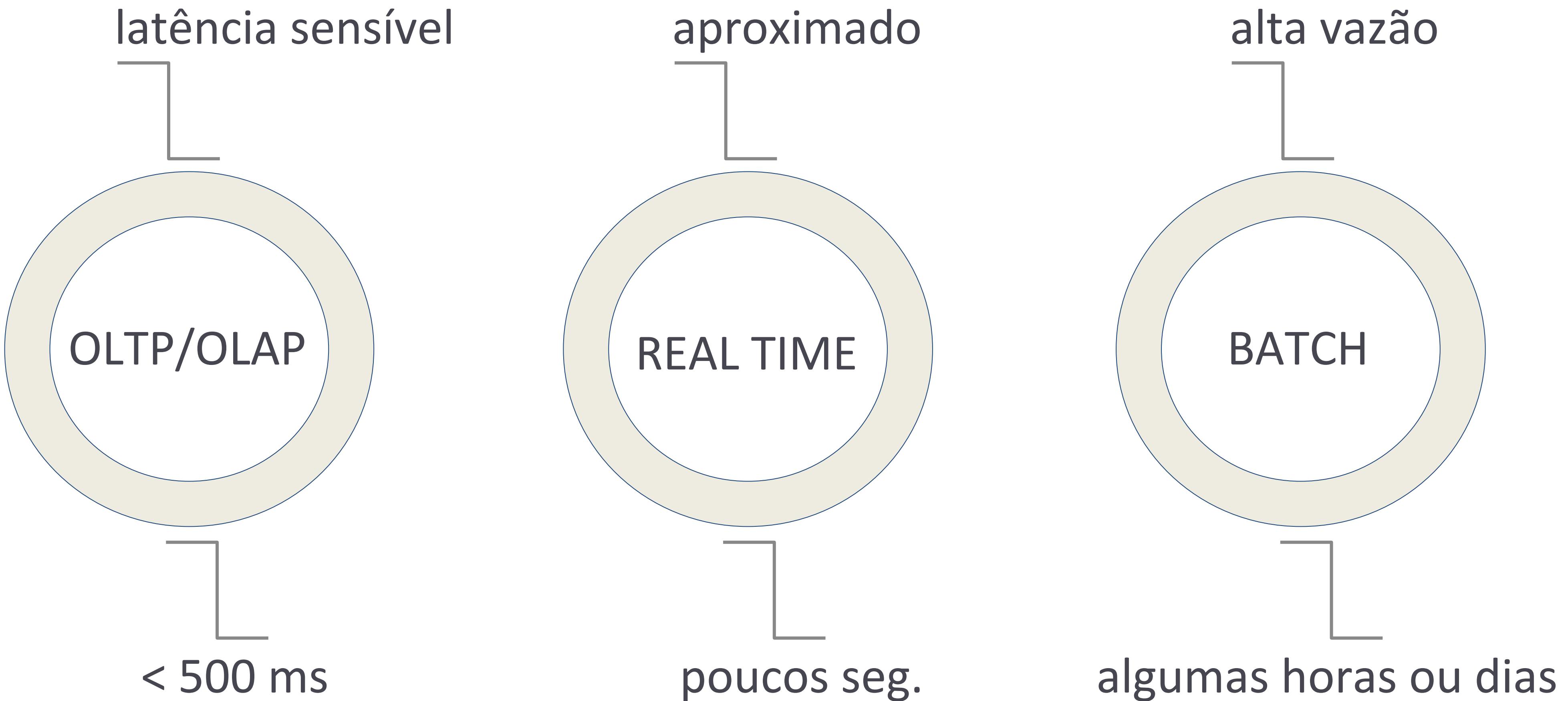


Deixar instalando...

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?



Apache Kafka

Apache Kafka

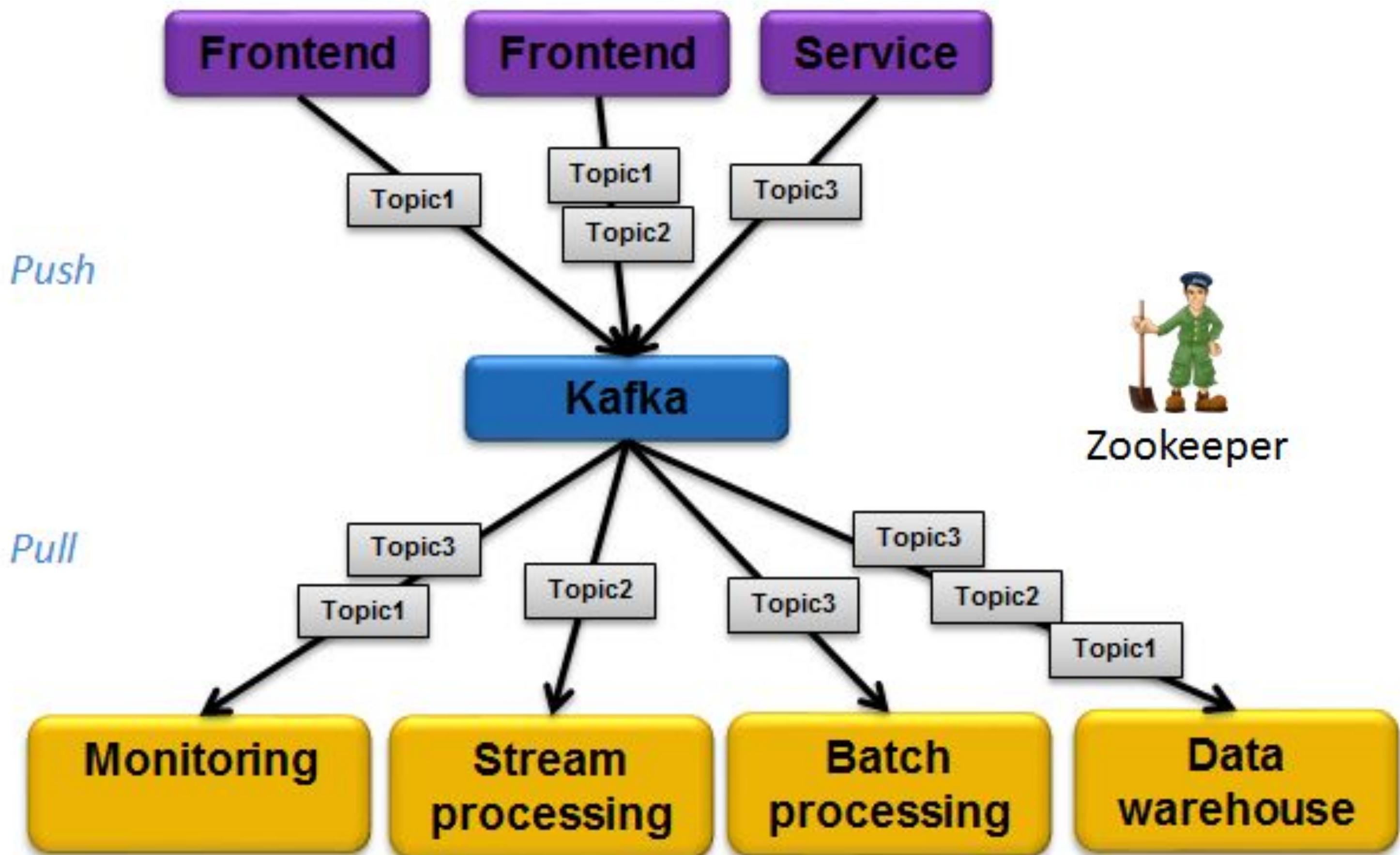
- Sistema de mensagens
 - Distribuído
 - Com alta vazão (*throughput*)
 - De geração (publicação) e leitura (sub-inscrição)
- Principais casos de uso:
 - Agregação de log
 - Processamento em tempo real
 - Monitoramento

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.

Kafka: conceitos

Producers



Broker

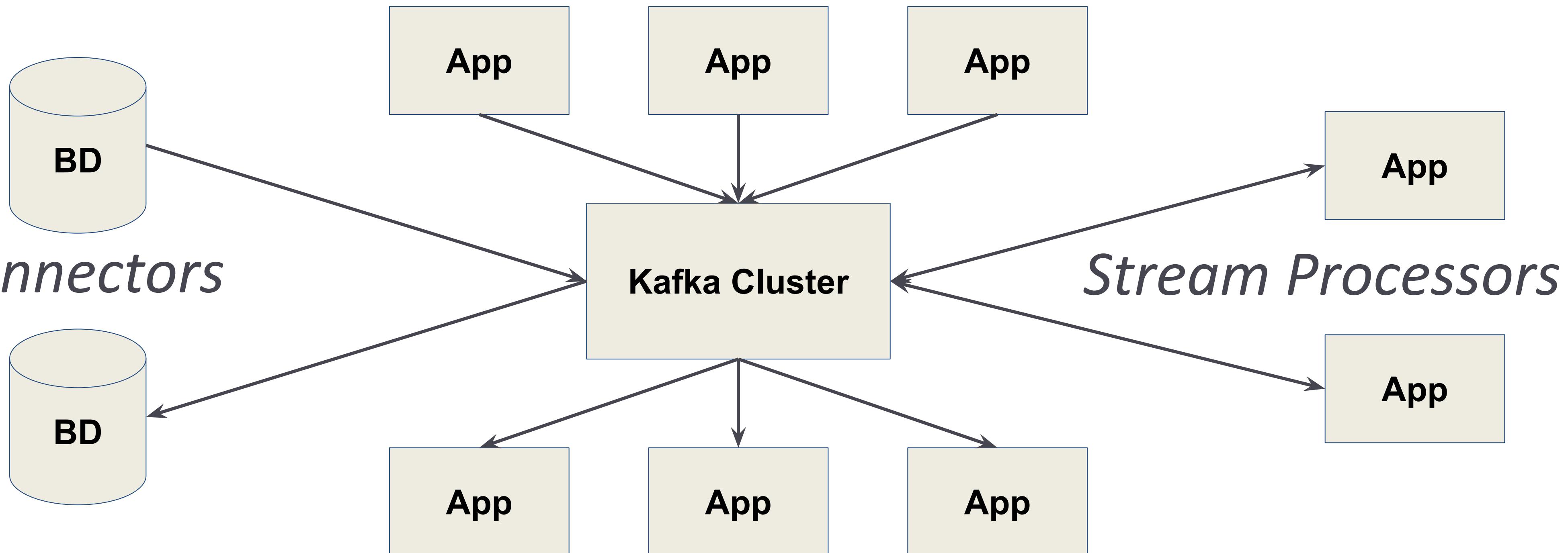
Consumers



Zookeeper

Kafka: arquitetura

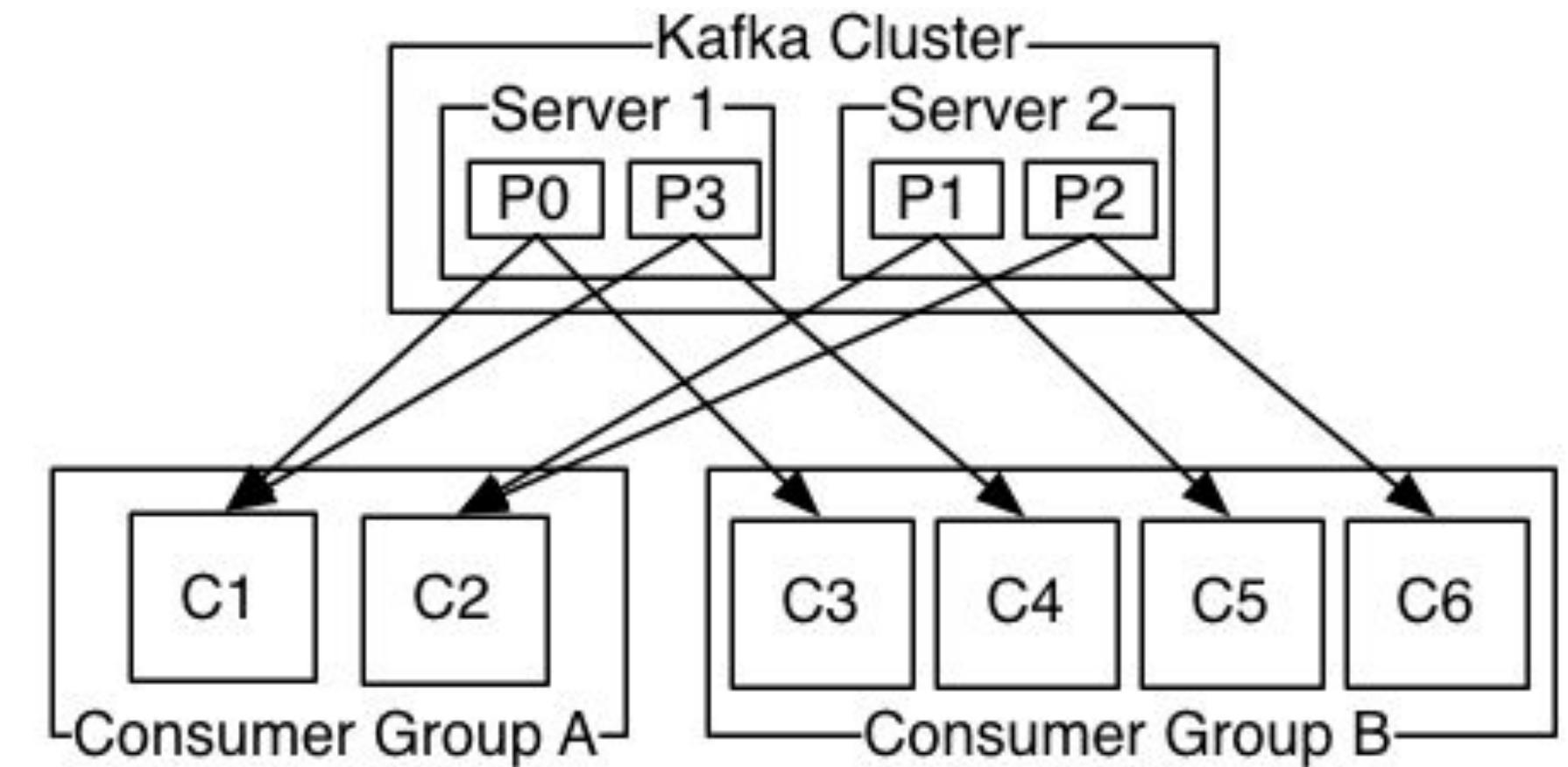
Producers



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 ($>>10000$).
- Não otimizado para latências de milissegundos.

Finalizando
nossa Setup...

Instalando algumas libs

Instalar o curl

```
➤ sudo apt install curl
```

Instalar o VS Code

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

Instalando algumas libs

Instalar o java, git e scala

```
➤ sudo apt install default-jdk scala git -y
```

Verificar as versões das libs instaladas

```
➤ java -version; javac -version; scala -version;  
git --version
```

Dúvidas?