Streaming de Dados em Tempo Real: Aula 1

Prof. Felipe Timbó

UNI7

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

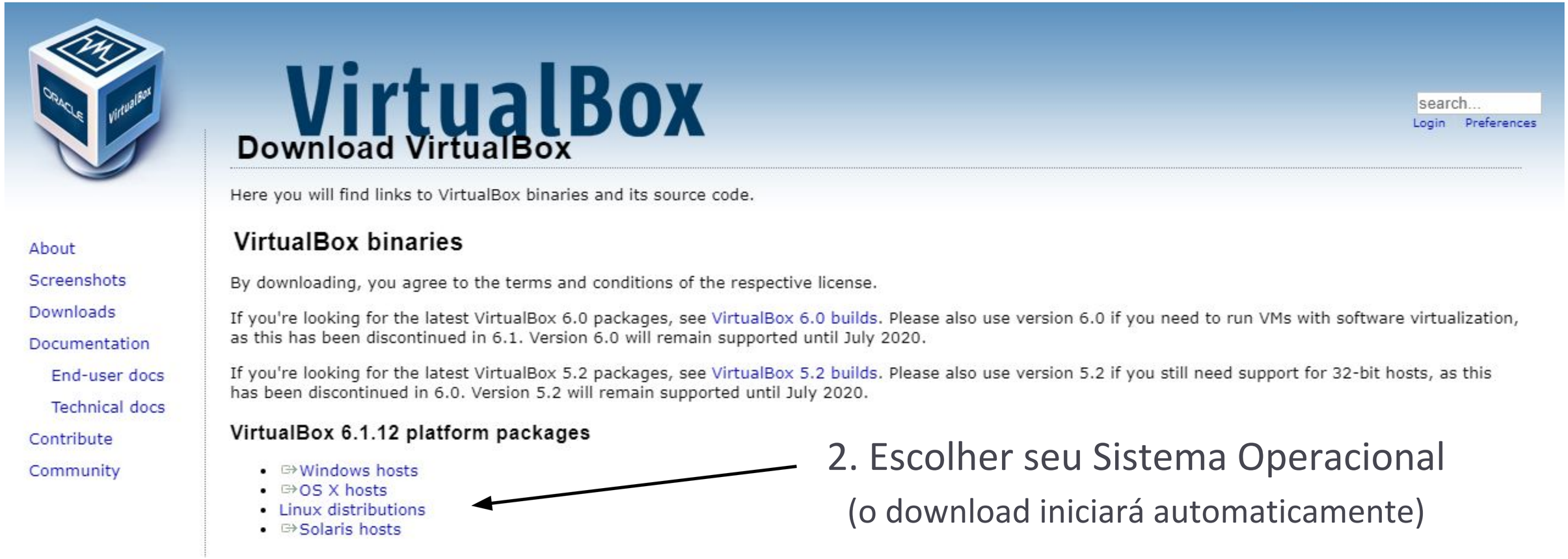
Arquivos no Microsoft Teams:

Slides, dados, scripts, livros e artigos

Antes de tudo...

Download do VirtualBox

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



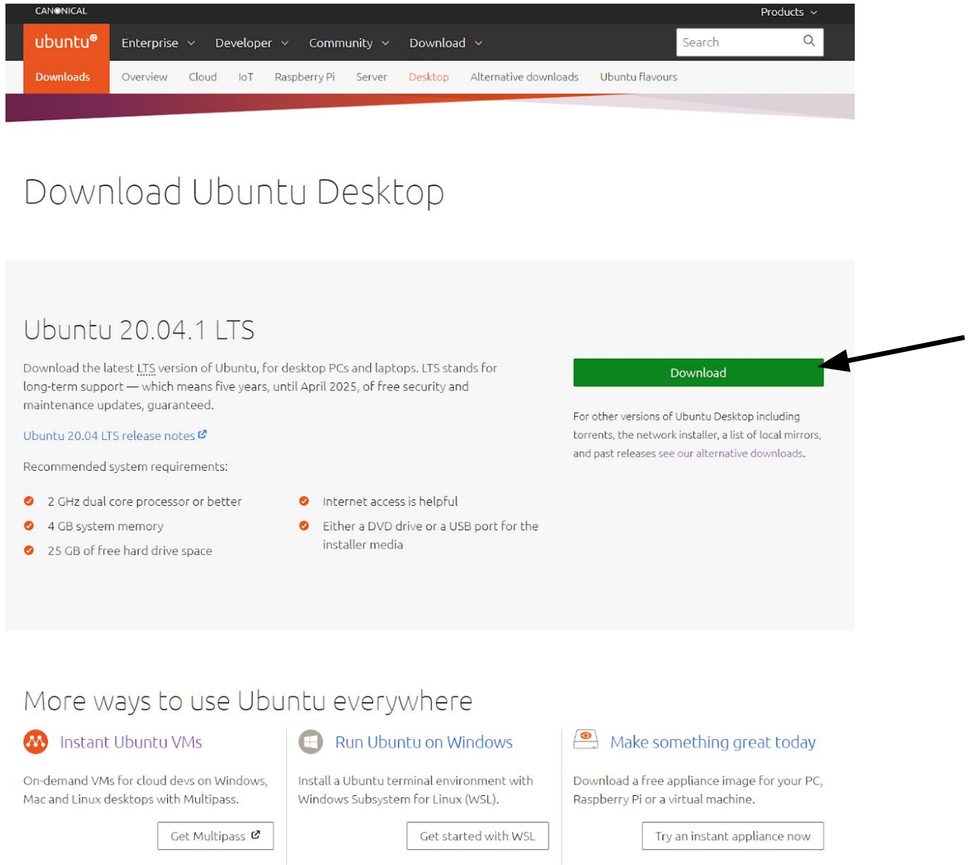
2. Escolher seu Sistema Operacional

(o download iniciará automaticamente)

Download do Ubuntu

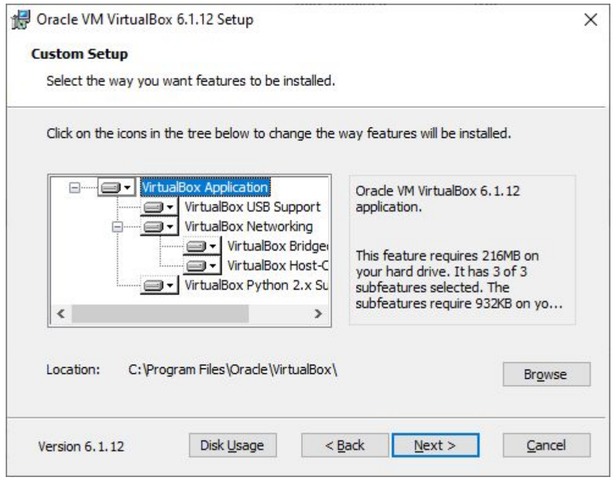
3. Acessar: https://ubuntu.com/download/desktop

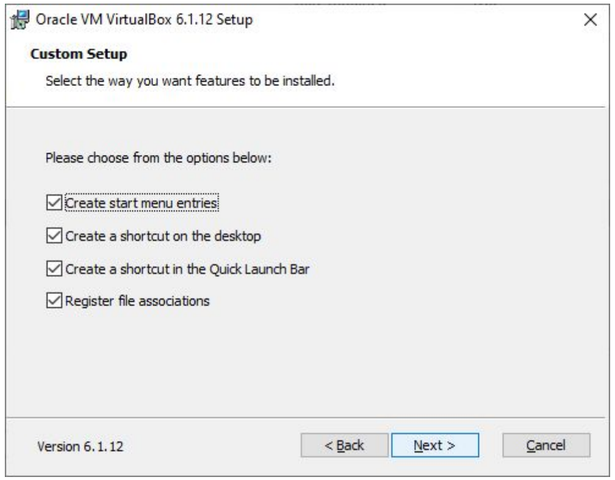
4. Realizar o download do Ubuntu

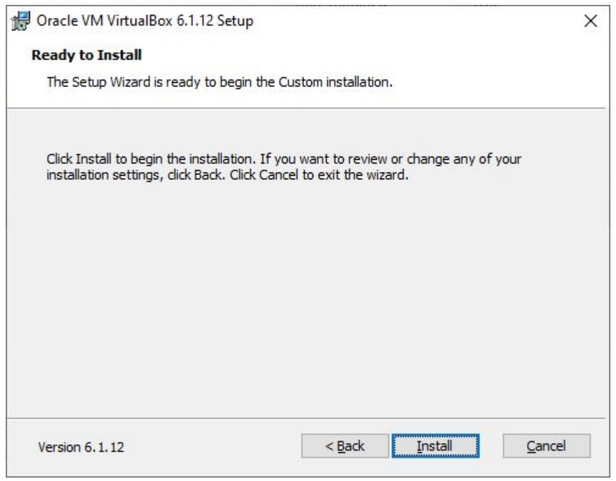


Instalação do VirtualBox

5. Instalar o VirtualBox

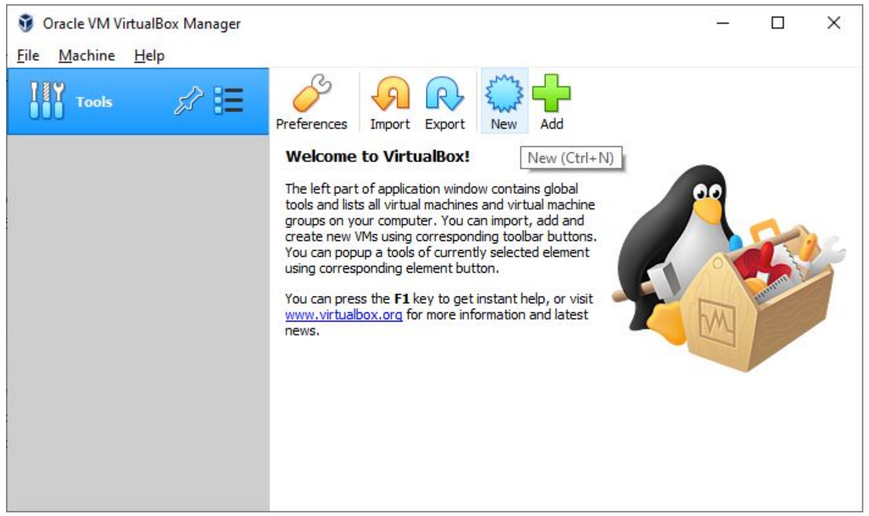


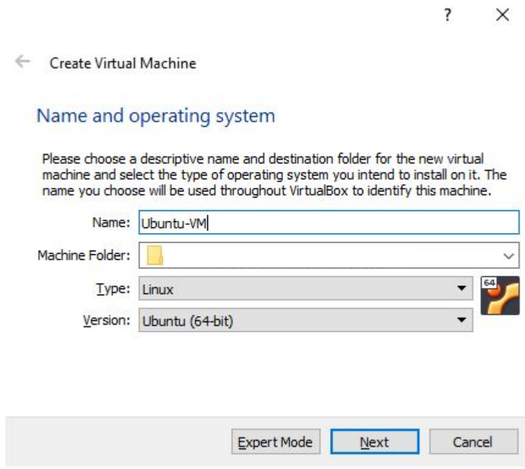


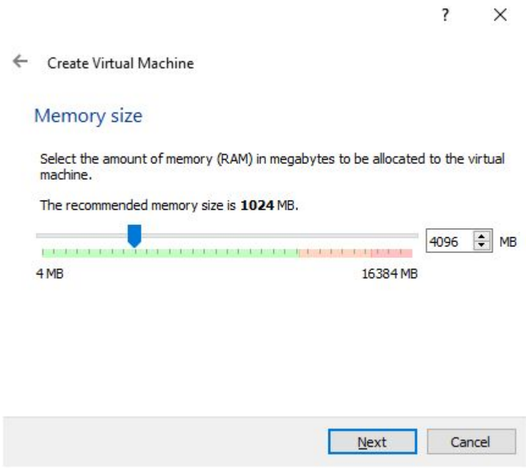


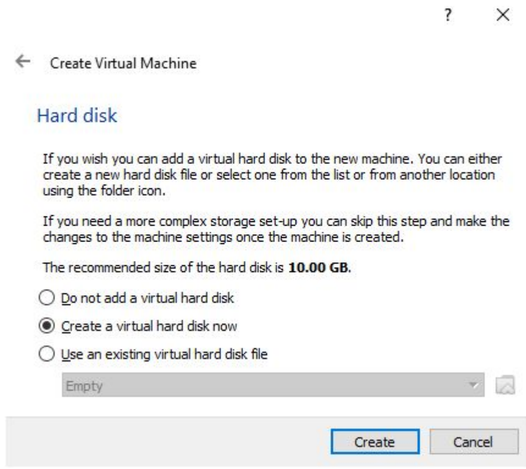
Criação de uma VM (Virtual Machine)

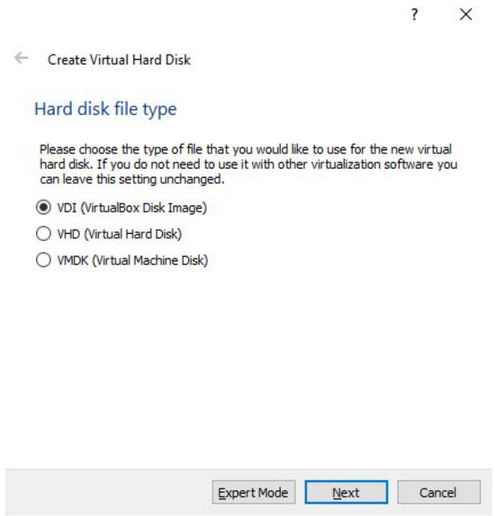
6. Criar uma nova Máquina Virtual Ubuntu

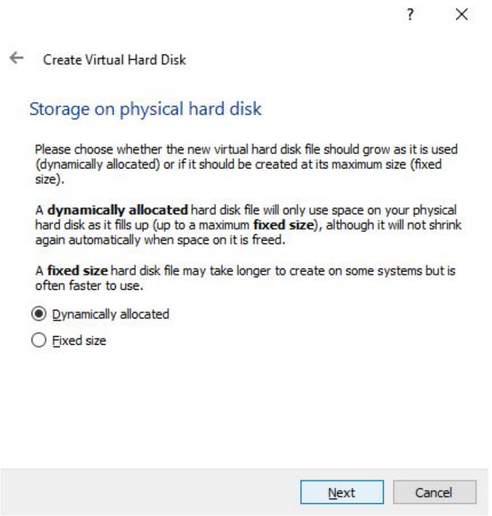


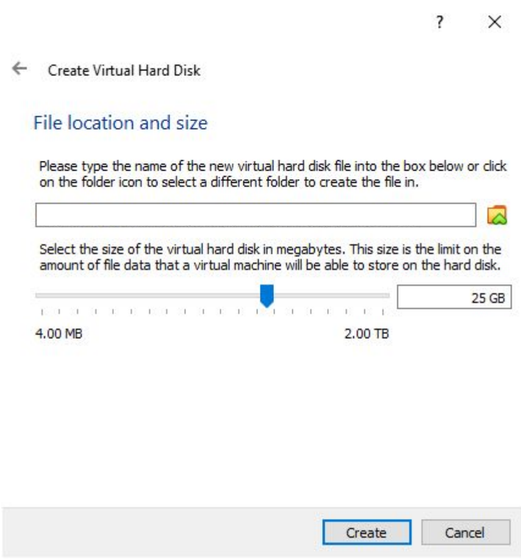


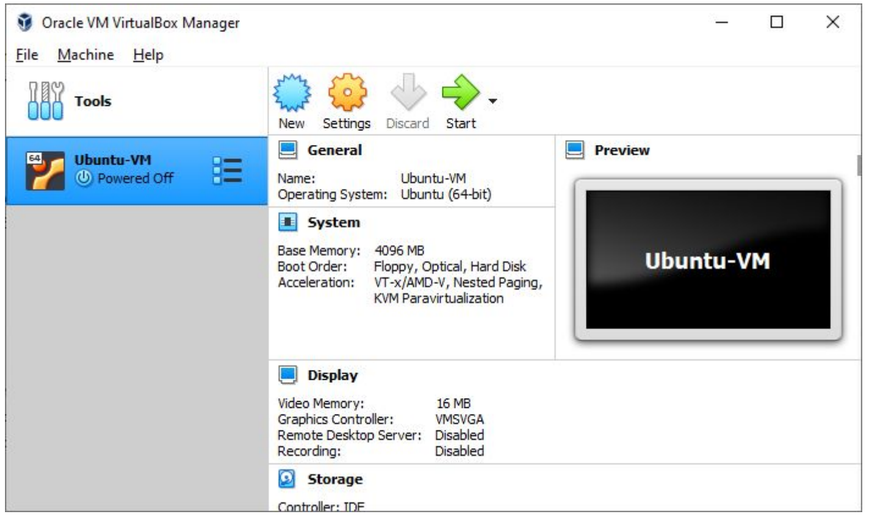






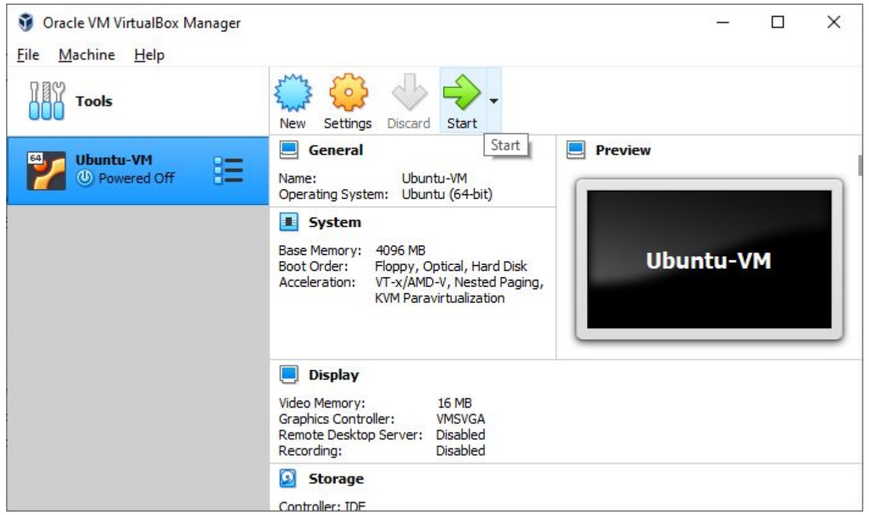


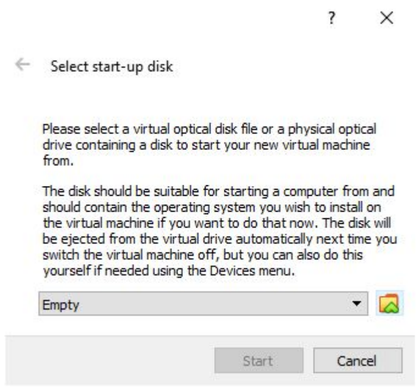


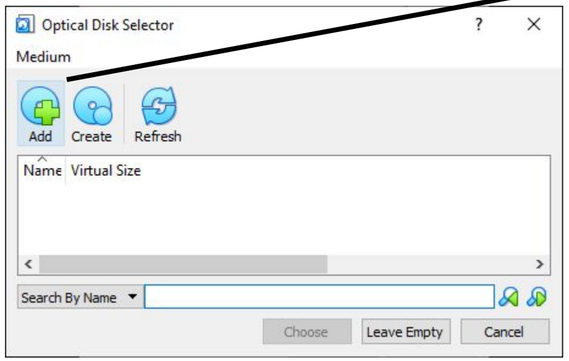


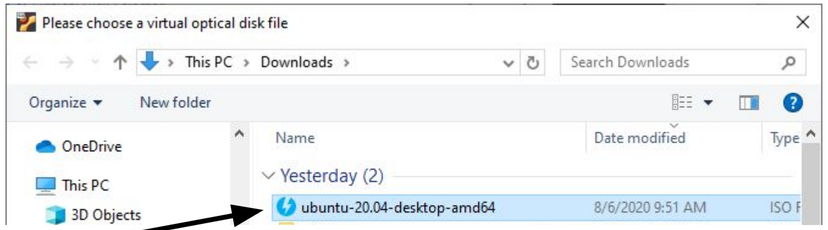
Criação de uma VM (Virtual Machine)

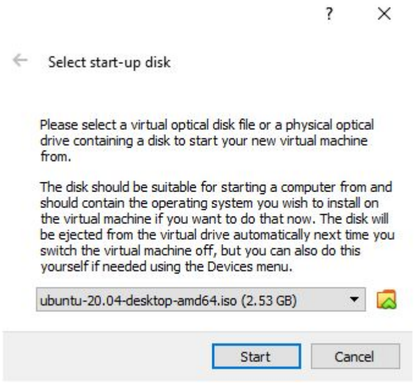
7. Iniciar a Máquina Virtual Ubuntu





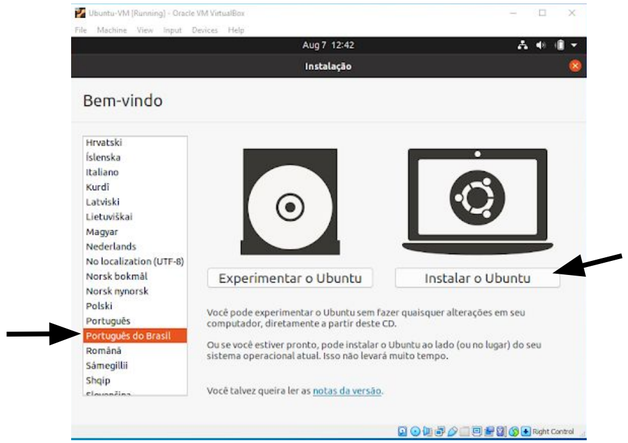




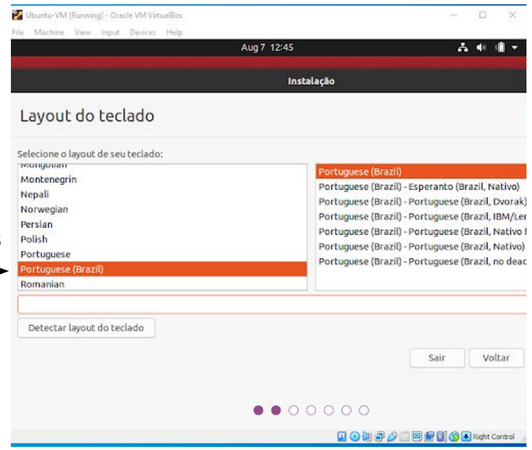


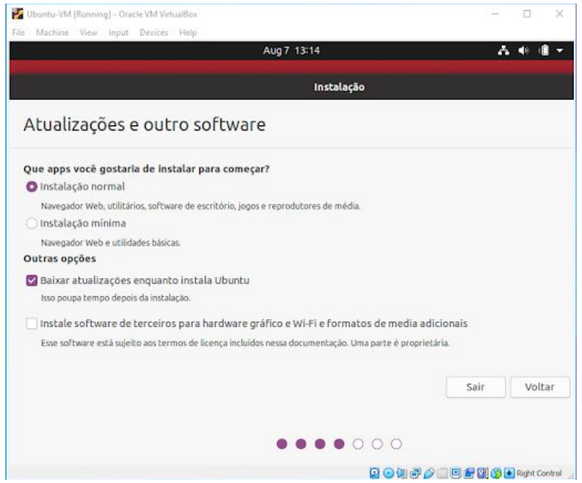
Instalação do Ubuntu

8. Instalar o Ubuntu na VM



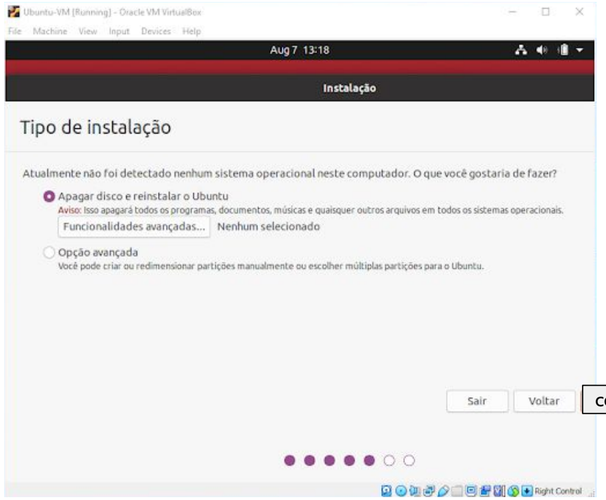
dois cliques





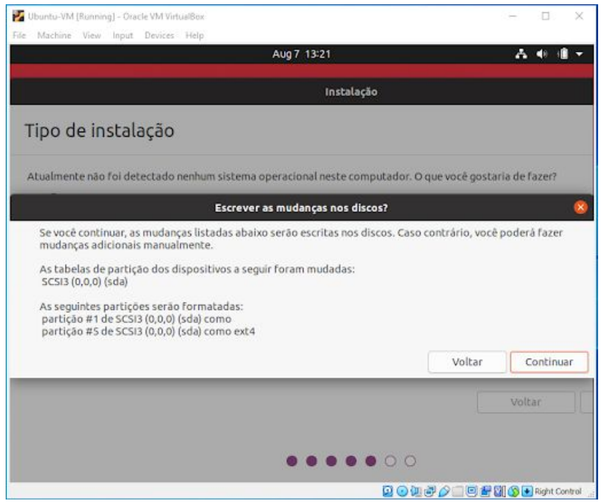
continuar

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



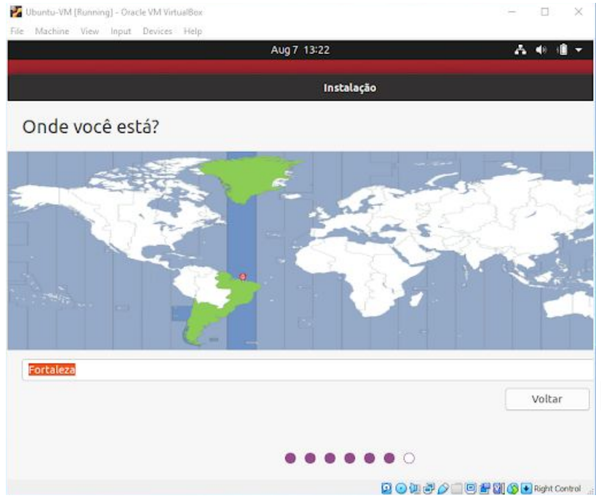
continuar

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



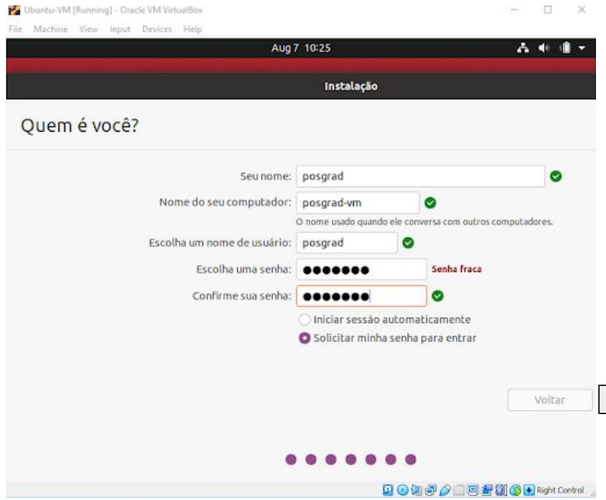
Instalação do Ubuntu

8. Instalar o Ubuntu na VM



continuar

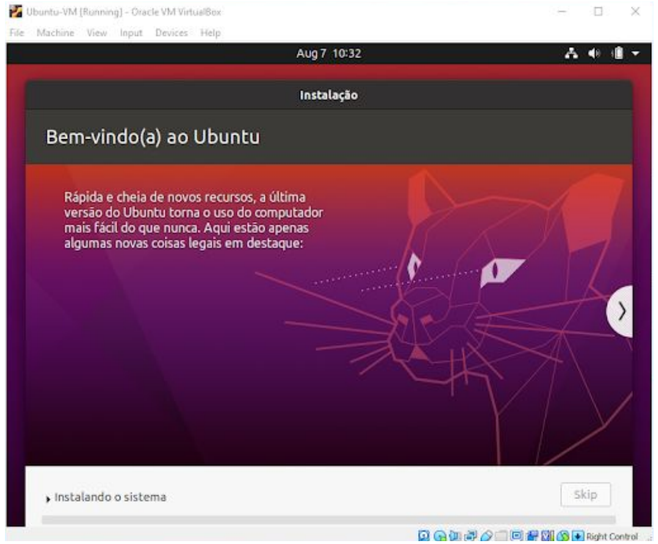
o botão ‘continuar’ pode estar escondido aqui. Acessá-lo via tecla ‘tab’



{ usuário: posgrad
senha: posgrad

continuar

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



Deixar instalando...

Introdução

O que acontece em 60 segundos?

Upload de 500h de vídeo no YouTube

54 milhões de mensagens no Whatsapp

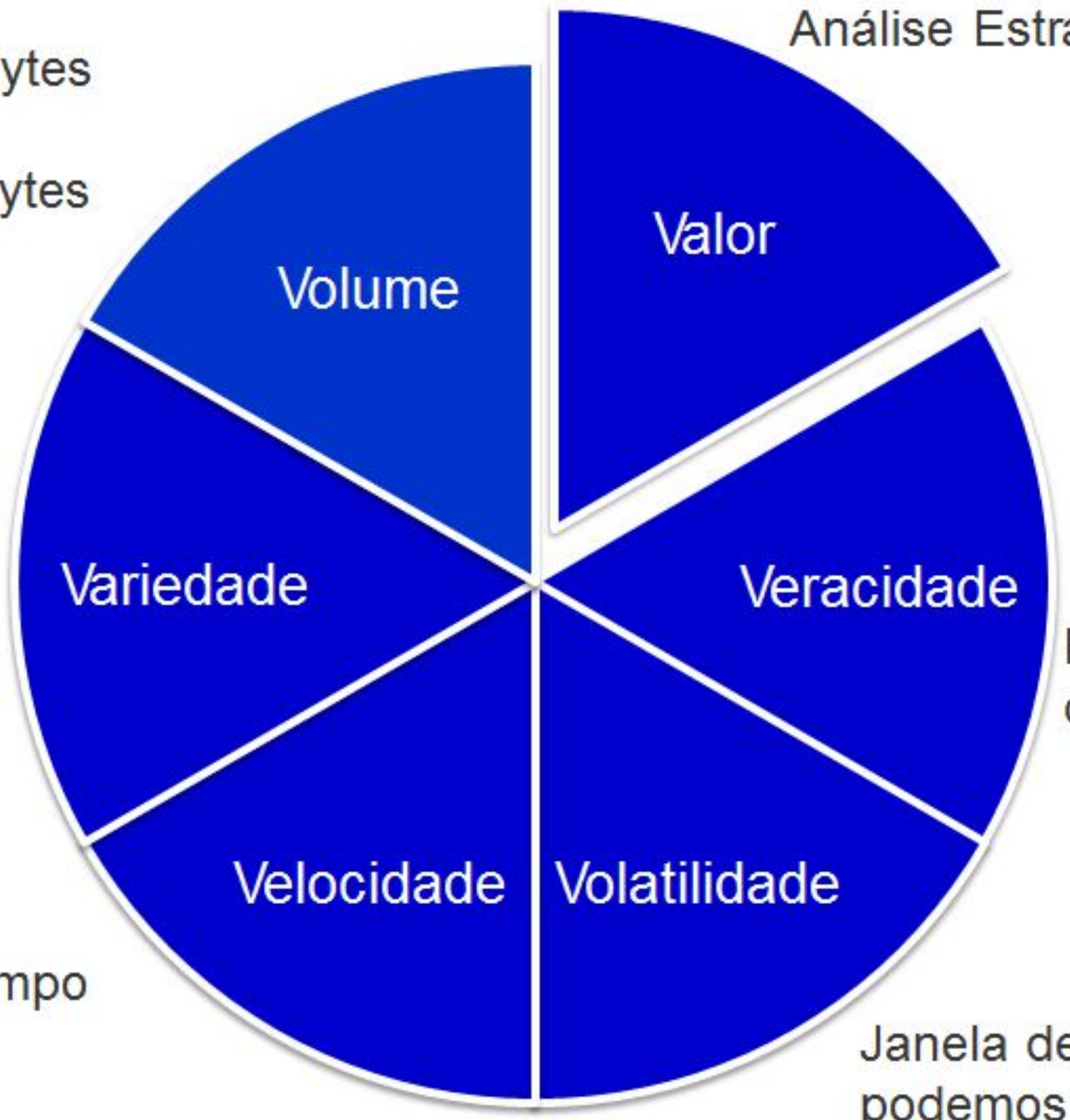
https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/

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



Terabytes
…
Exabytes

Análise Estratégica de Dados

Não-estruturado
Semi-estruturado
Estruturado

Estado verdadeiro
da realidade

Batch Tempo
Real

Janela de tempo onde
podemos usar os dados

Processamento em Batch

Todas as entradas

Hadoop,

Spark, etc.

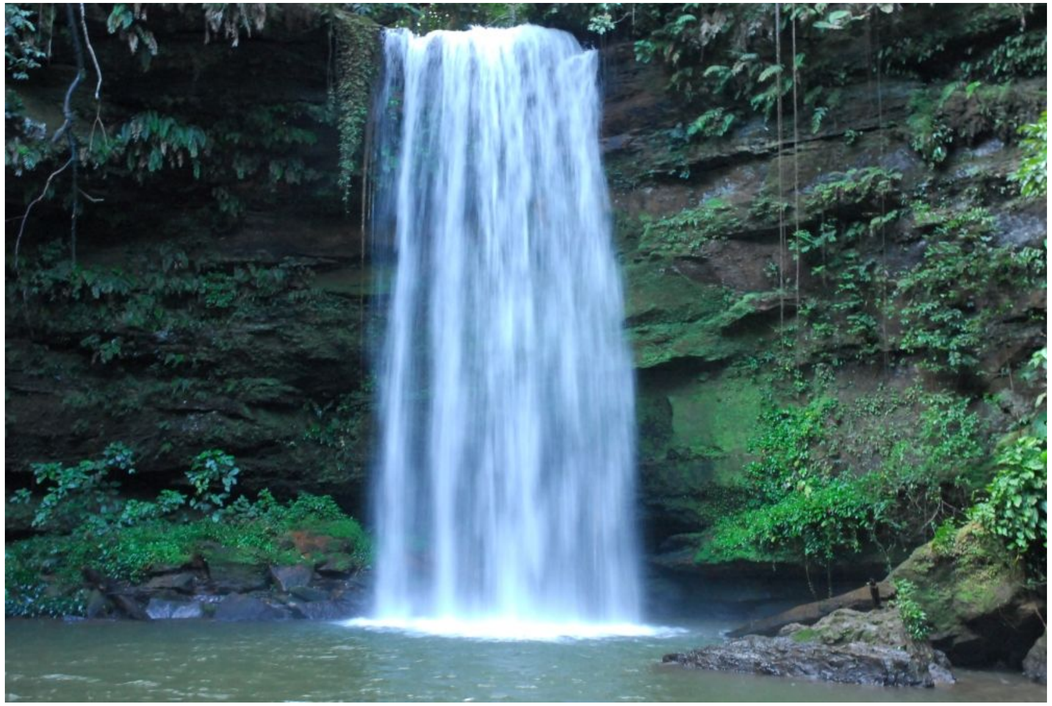
Todas as saídas

Streaming

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

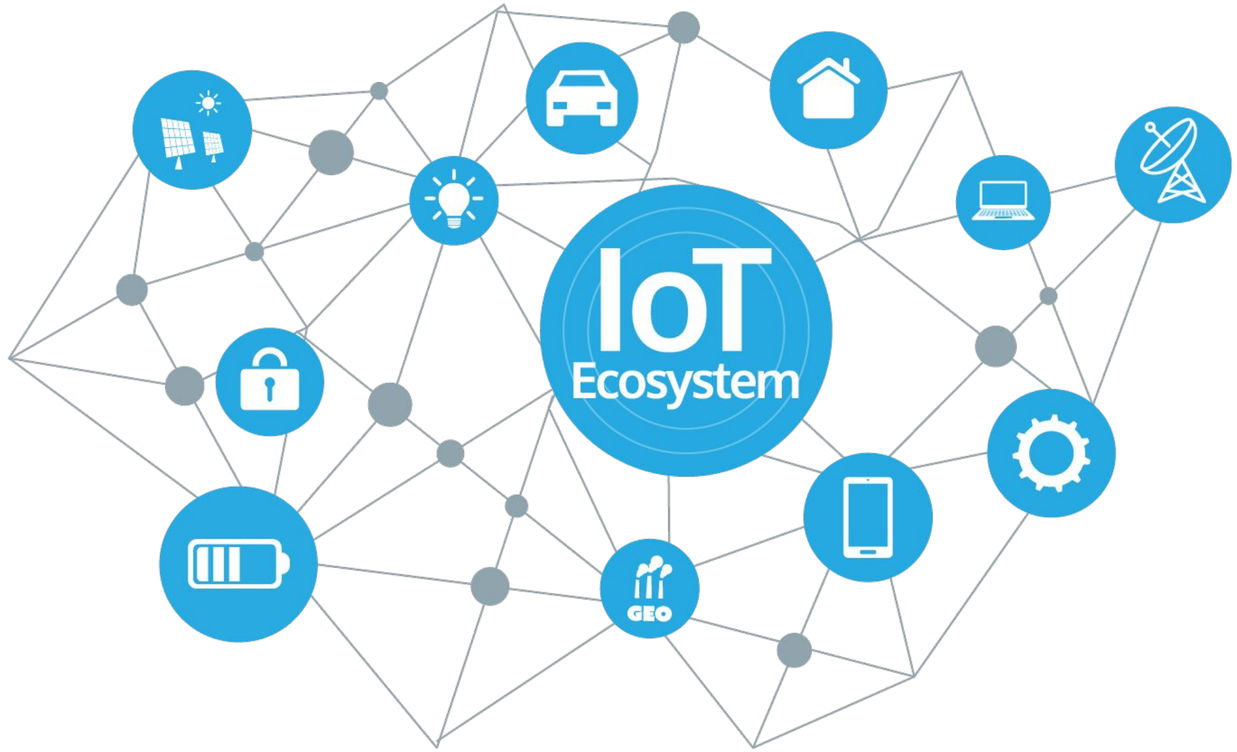
Streaming

Fluxo contínuo (contínuo ≠ constante).



Streaming de dados

Fluxo contínuo de dados.



Spotify®

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

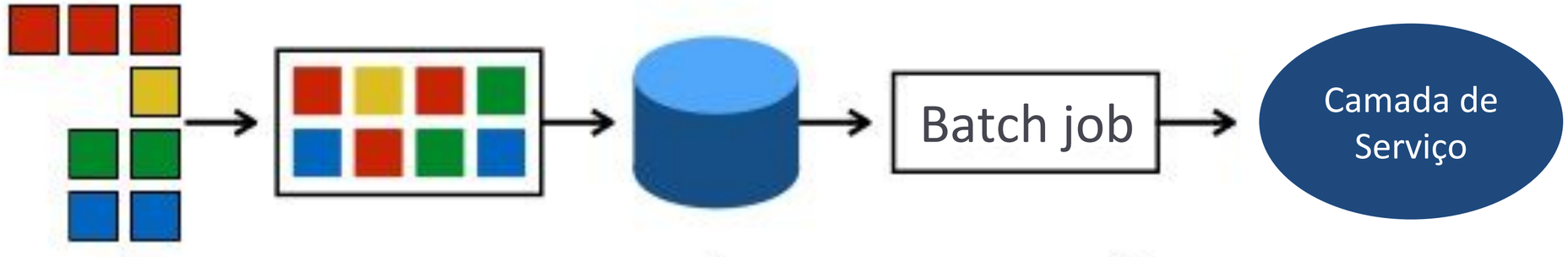
Dados em grafo: redes sociais

Dados de mobilidade: coordenadas geográficas x tempo

Etc.

Processamento de Streamings

Em geral:



continuamente produzido

arquivos são **streams finitos**

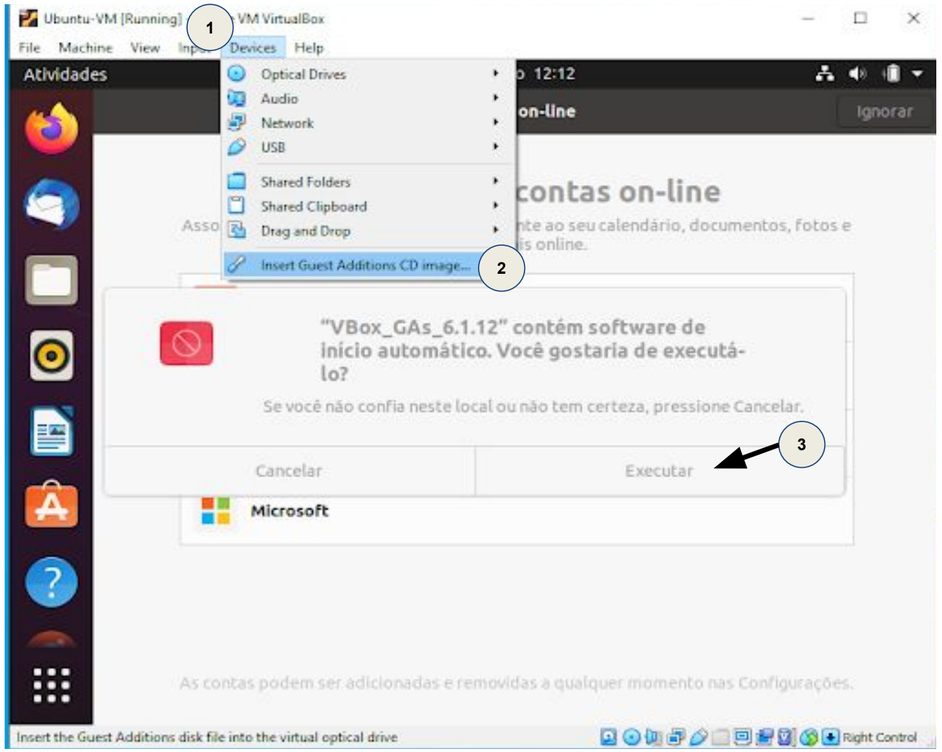
periodicamente executado

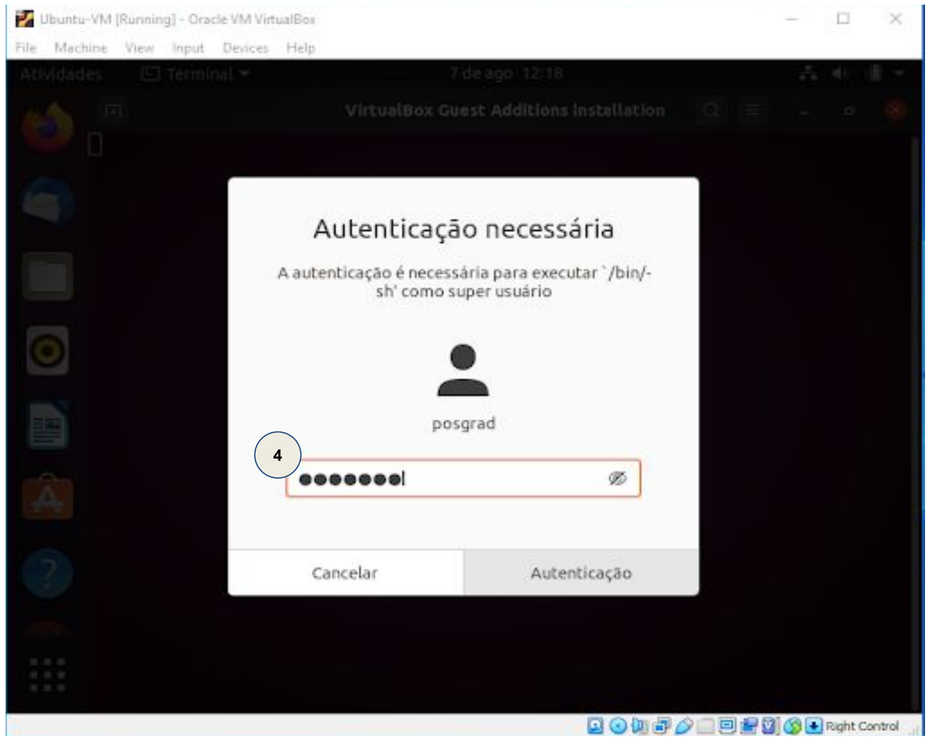
Continuando nosso Setup...

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

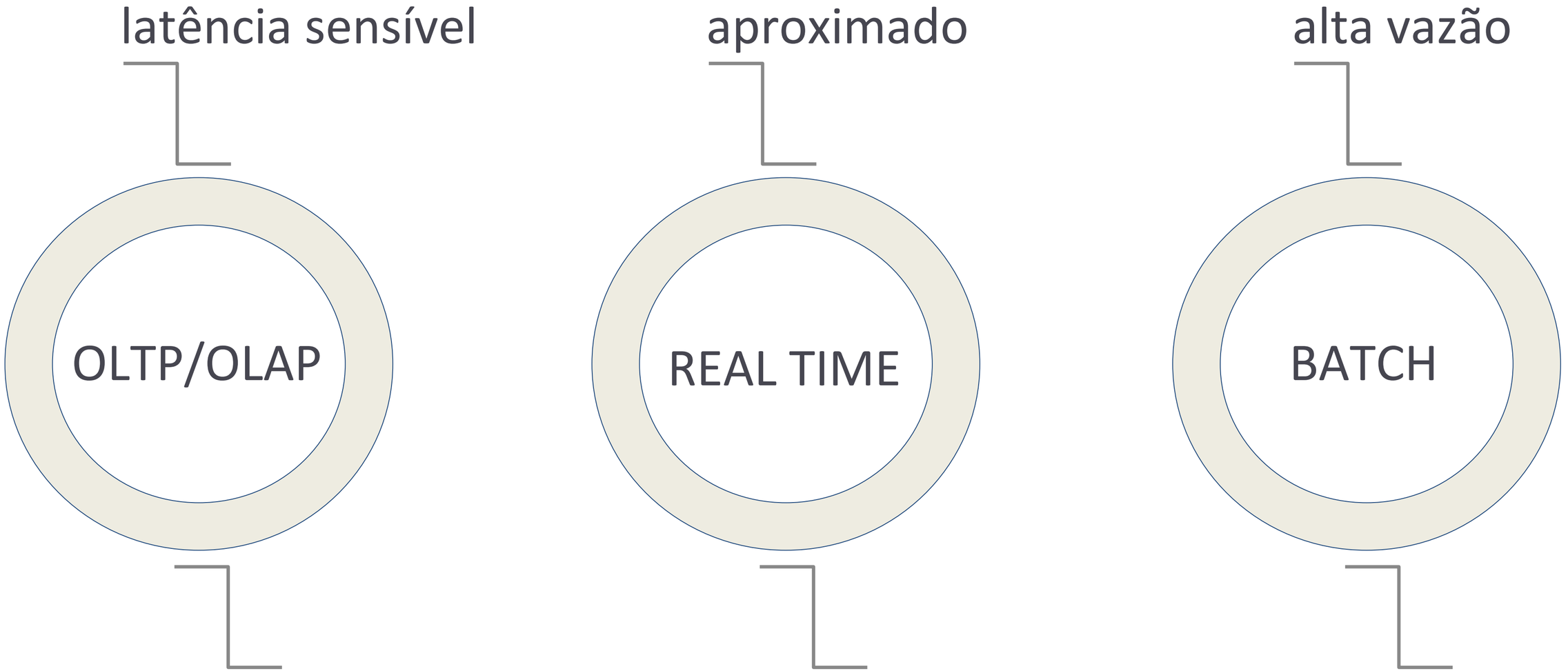
- `sudo apt update`
- `sudo apt upgrade`

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

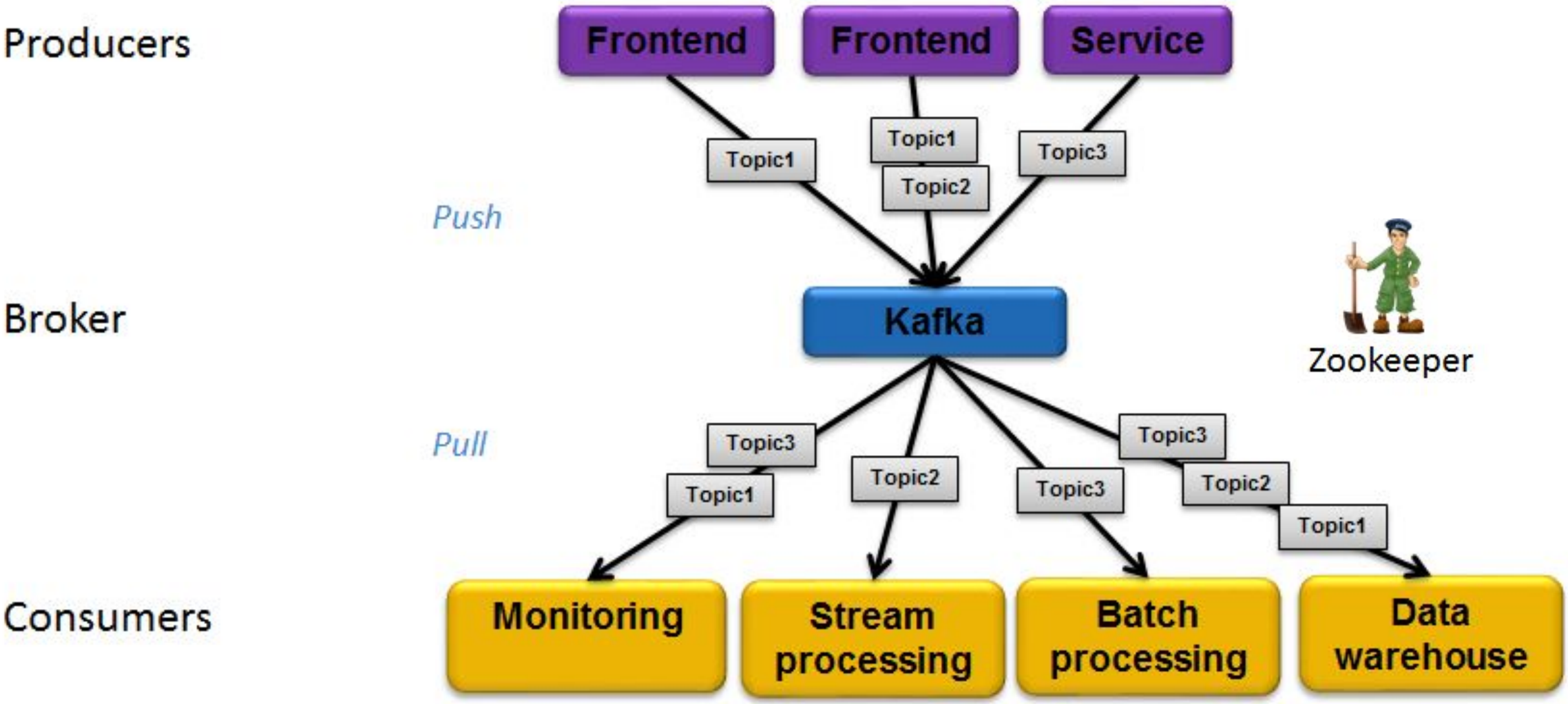
Apache Kafka

- 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

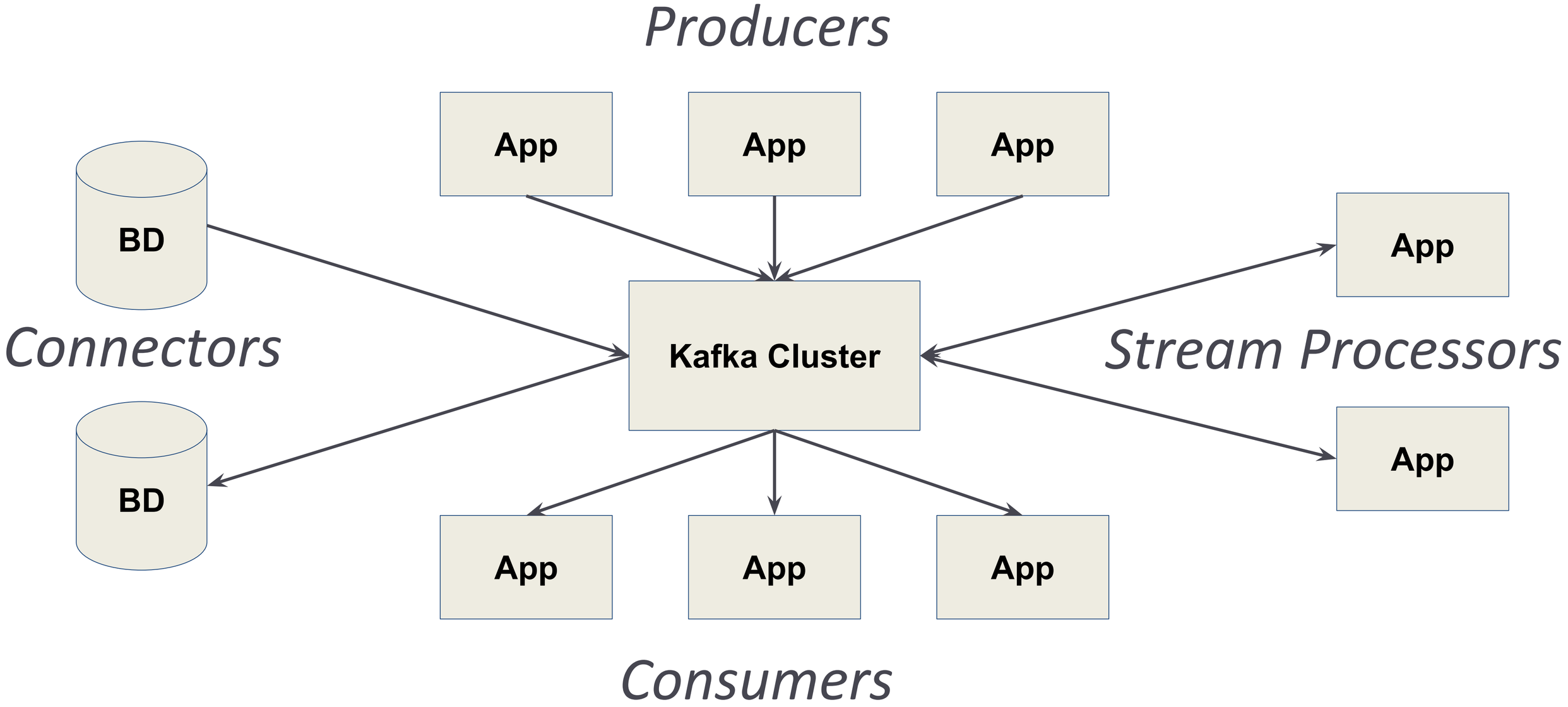
Apache Kafka

- 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

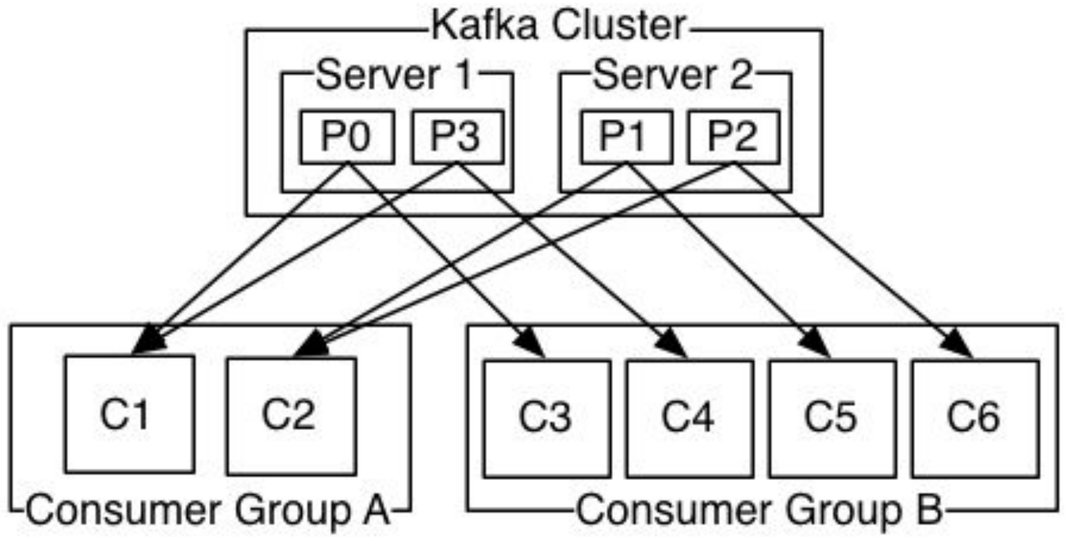


Kafka: arquitetura



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 nosso 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?