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 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

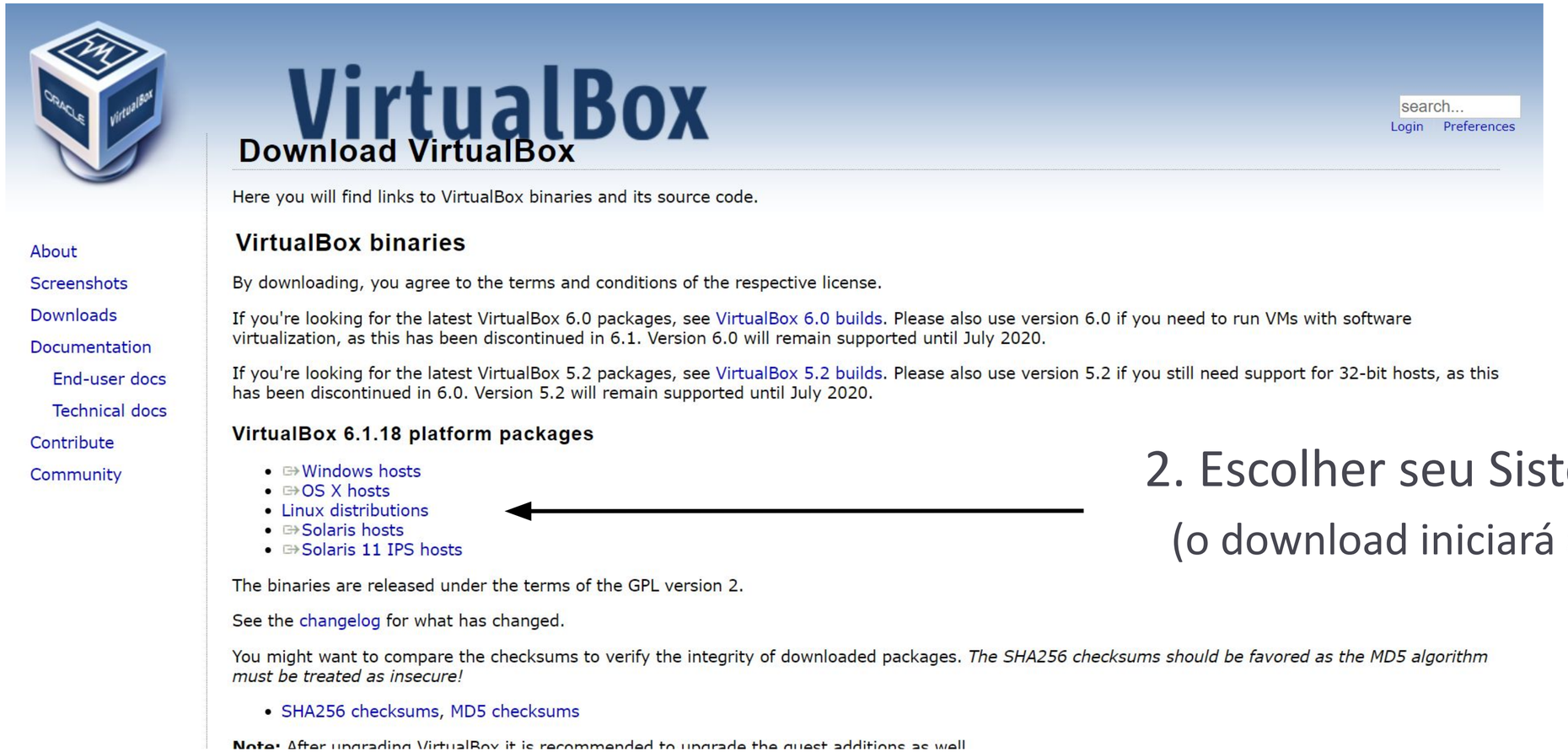
Arquivos (slides, livros e artigos)

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

Antes de tudo...

Download do VirtualBox

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



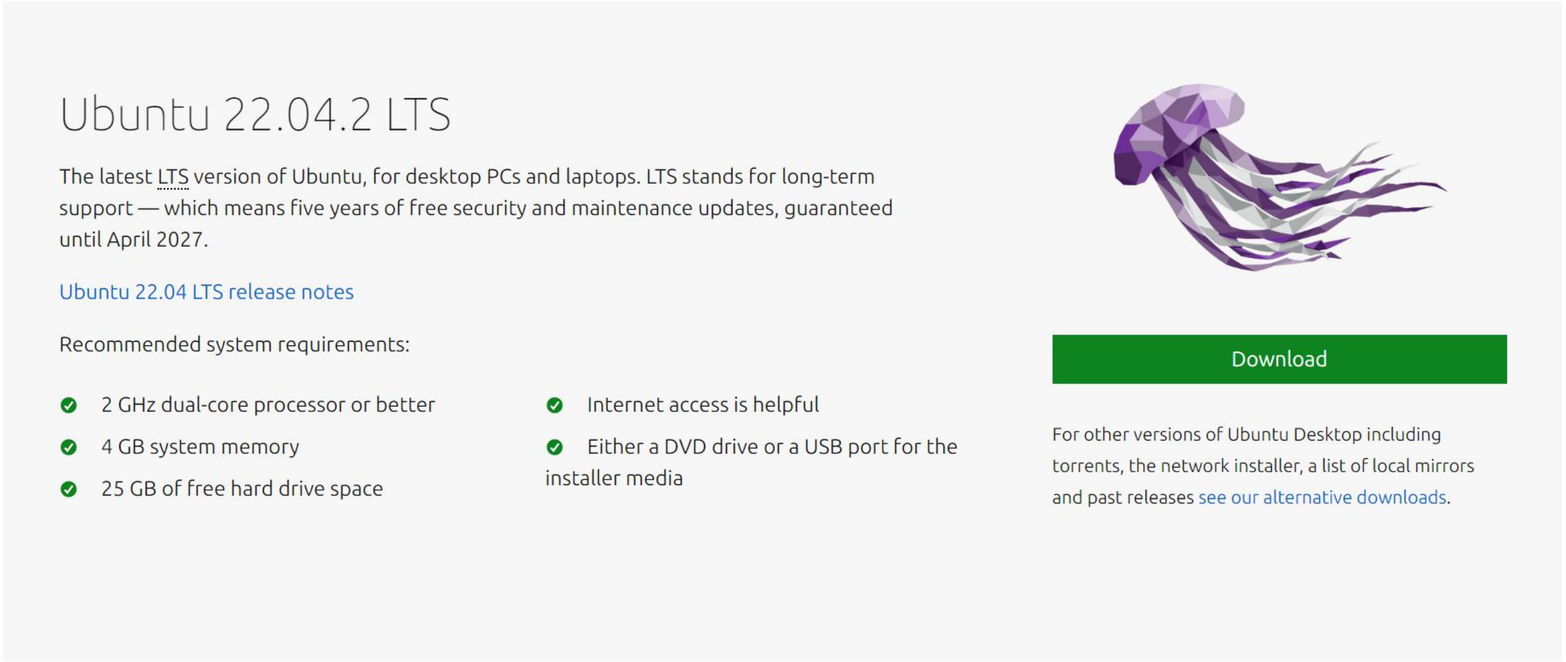
2. Escolher seu Sistema Operacional

(o download iniciará automaticamente)

Download do Ubuntu

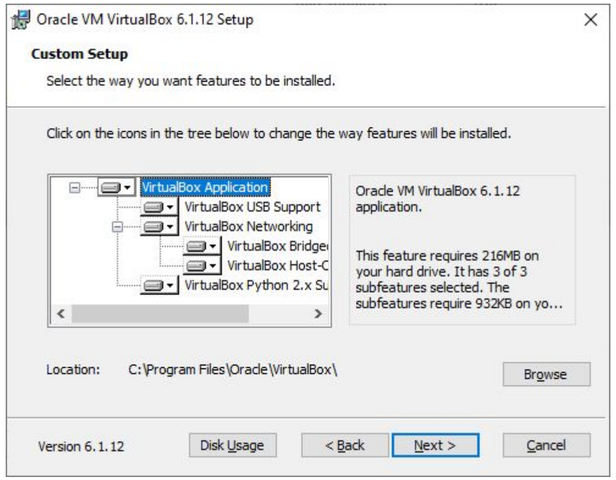
3. Acessar: https://ubuntu.com/download/desktop

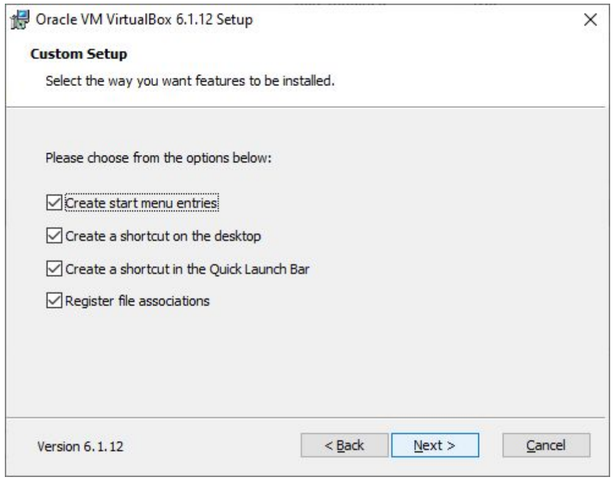
4. Realizar o download do Ubuntu 22.04.2 LTS

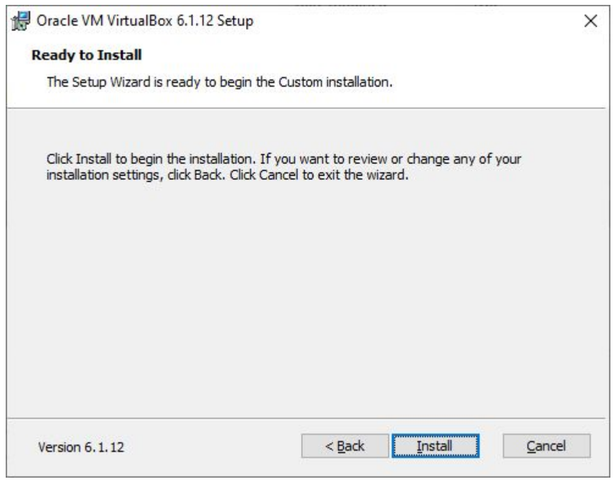


Instalação do VirtualBox

5. Instalar o VirtualBox

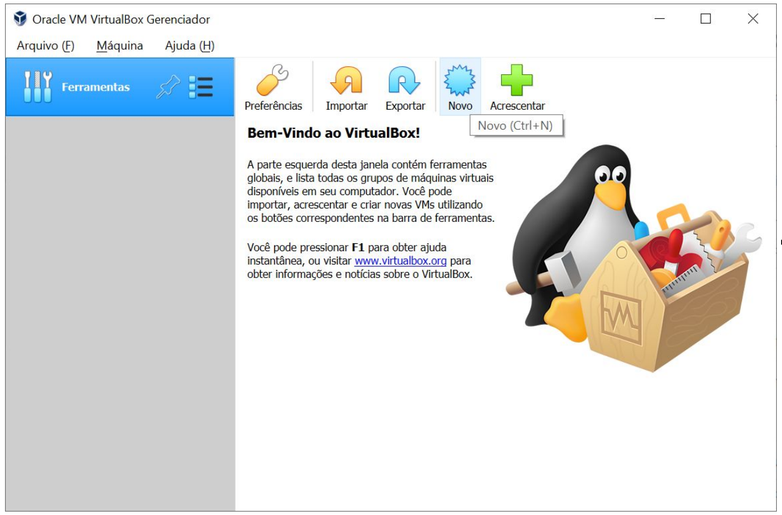


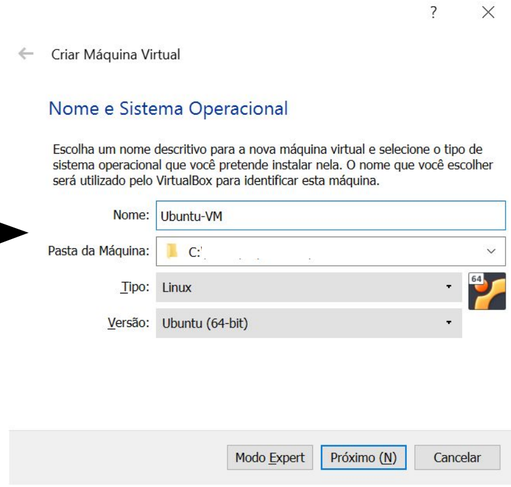


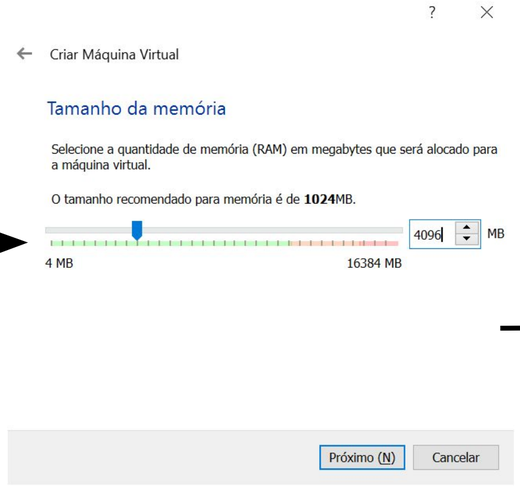


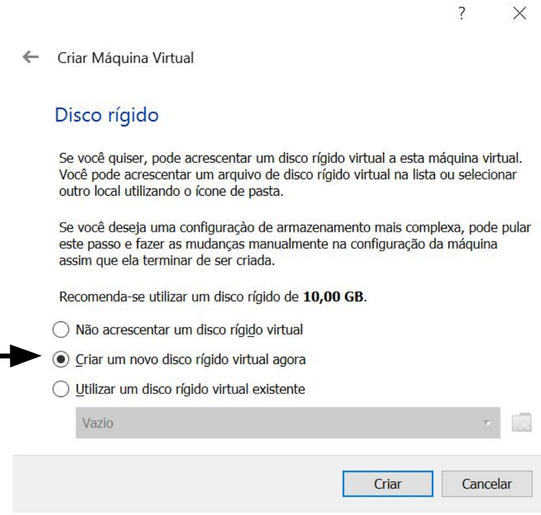
Criação de uma VM (Virtual Machine)

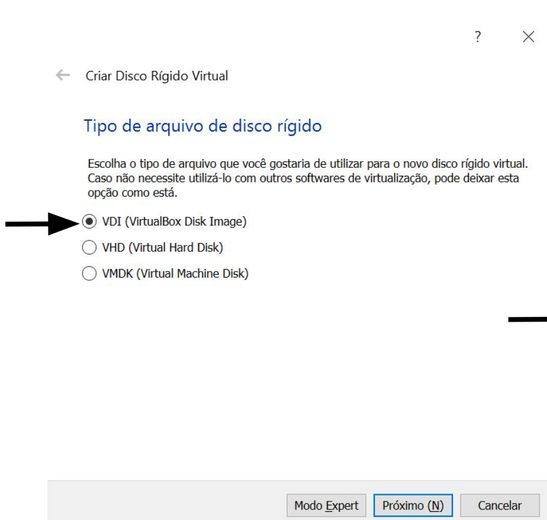
6. Criar uma nova Máquina Virtual Ubuntu

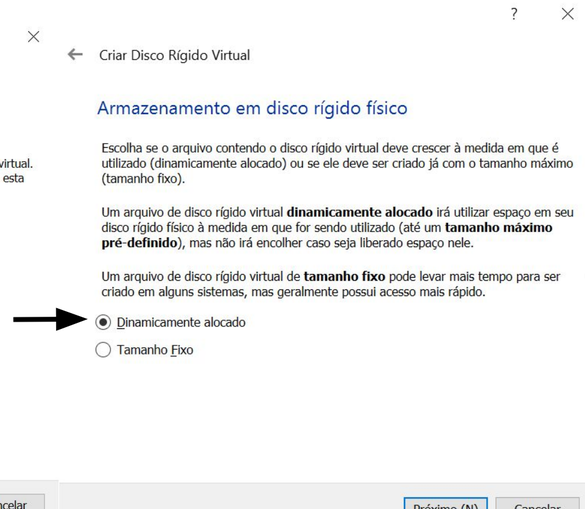


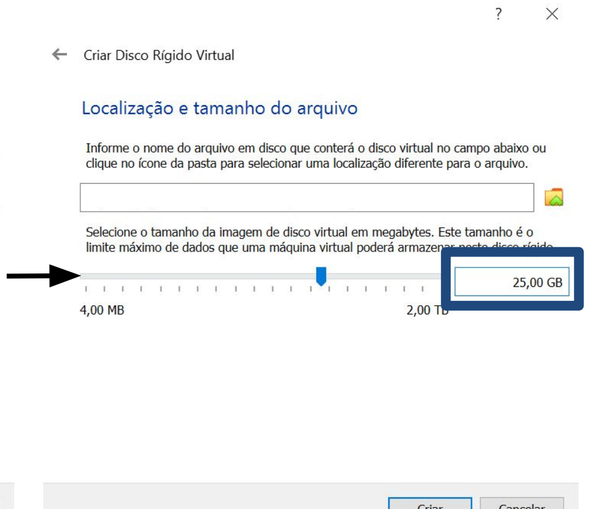


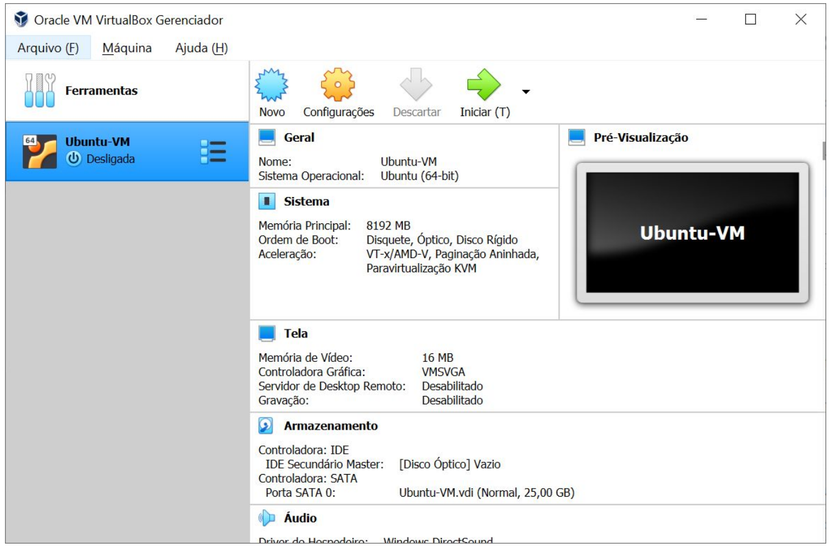






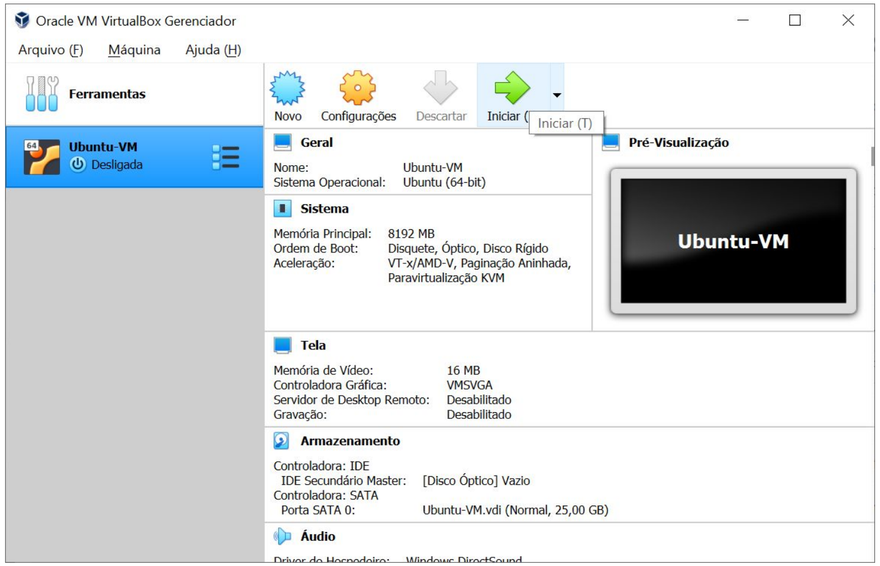


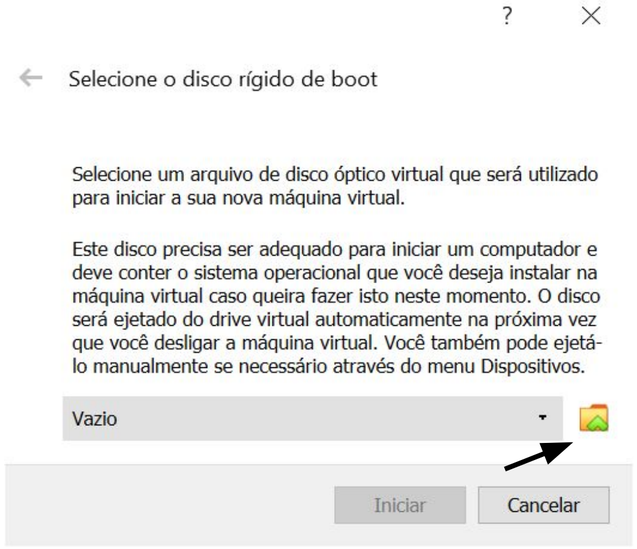


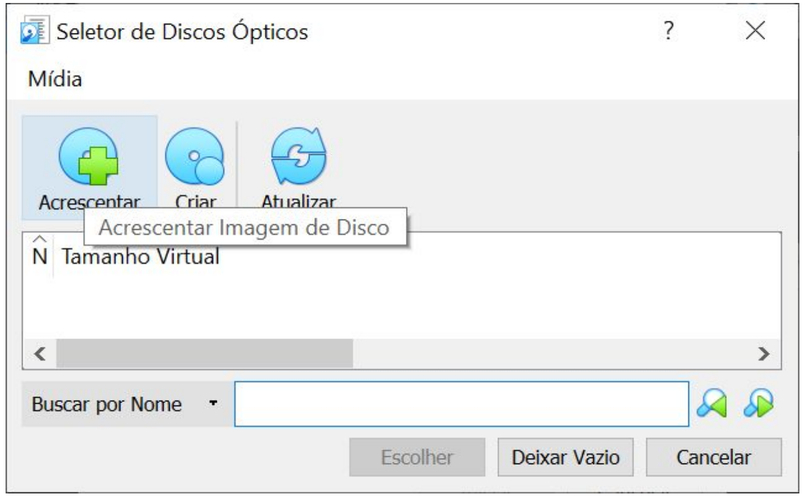


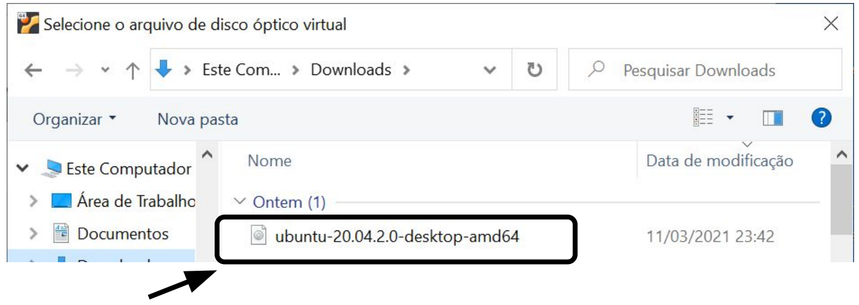
Criação de uma VM (Virtual Machine)

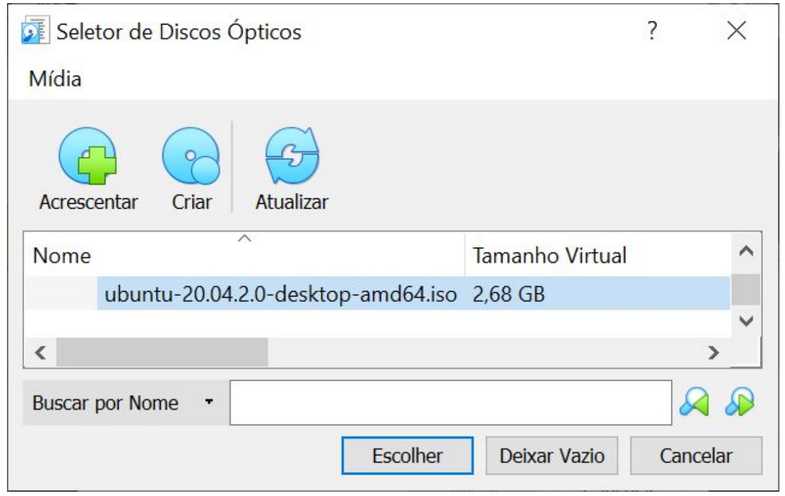
7. Iniciar a Máquina Virtual Ubuntu

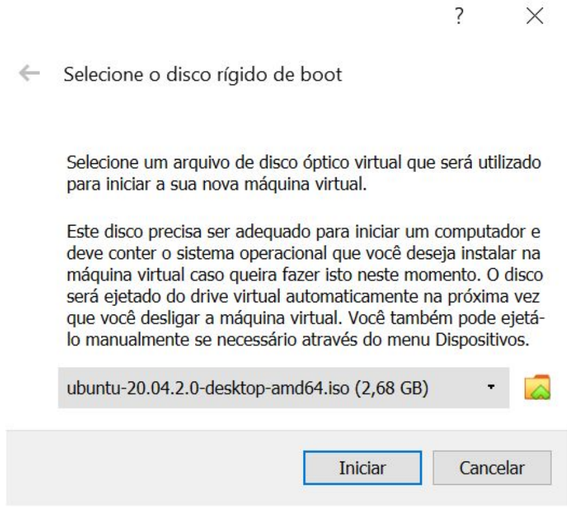






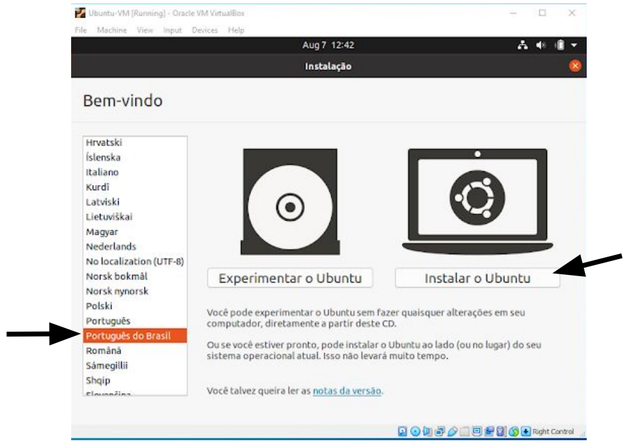




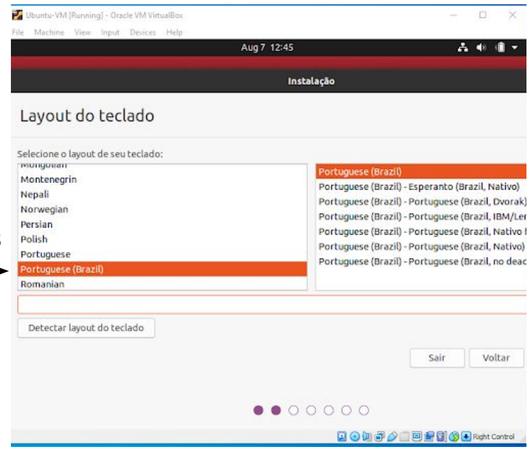


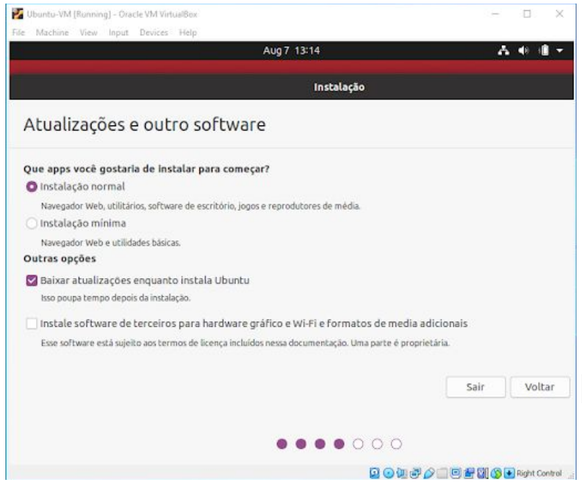
Instalação do Ubuntu

8. Instalar o Ubuntu na VM



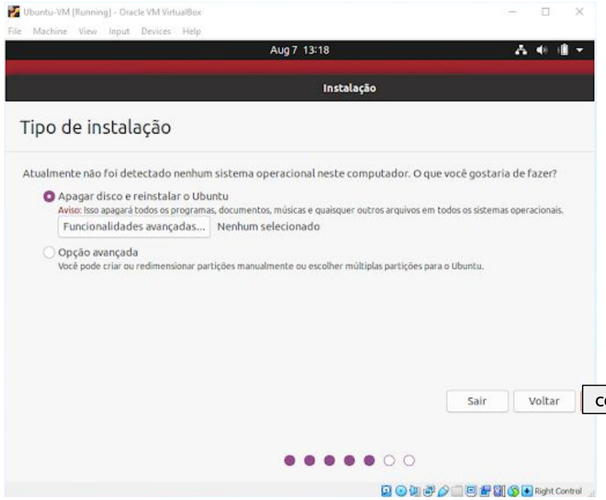
dois cliques





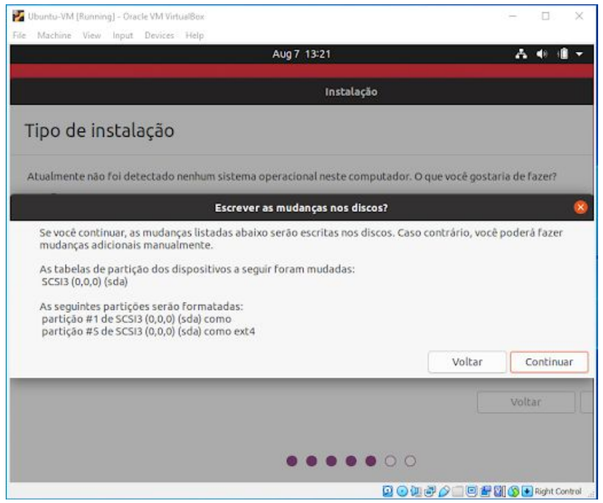
continuar

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



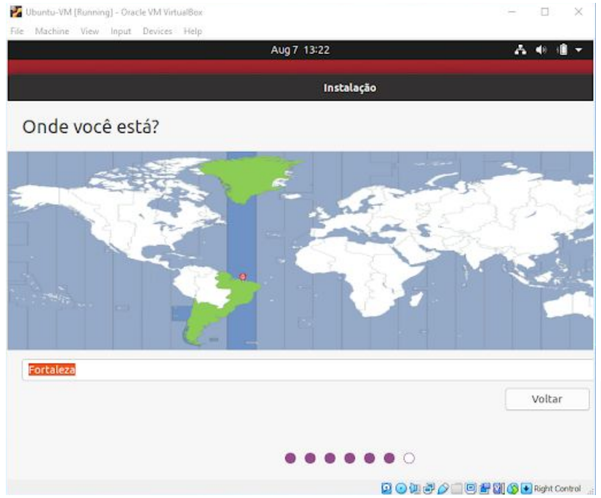
continuar

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



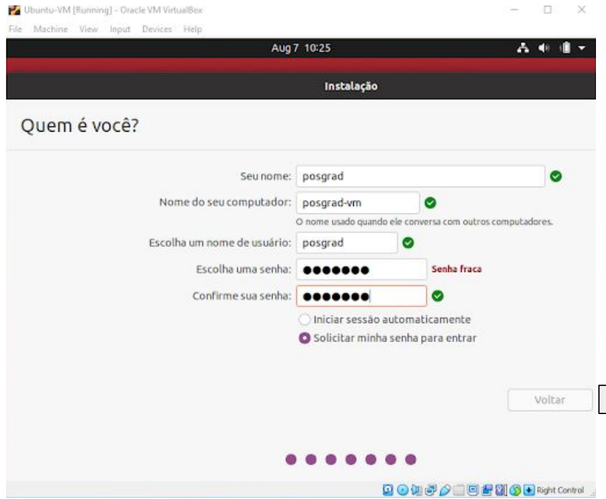
Instalação do Ubuntu

8. Instalar o Ubuntu na VM



continuar

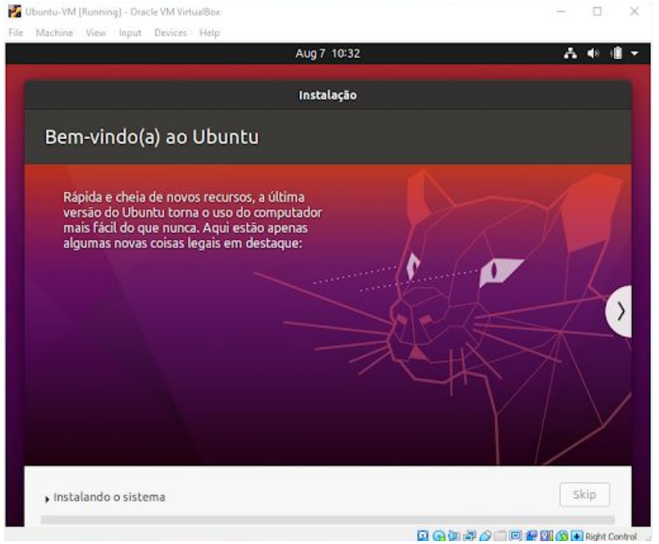
o botão ‘continuar’ pode estar escondido aqui. Acessá-lo via tecla ‘tab’



{ 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

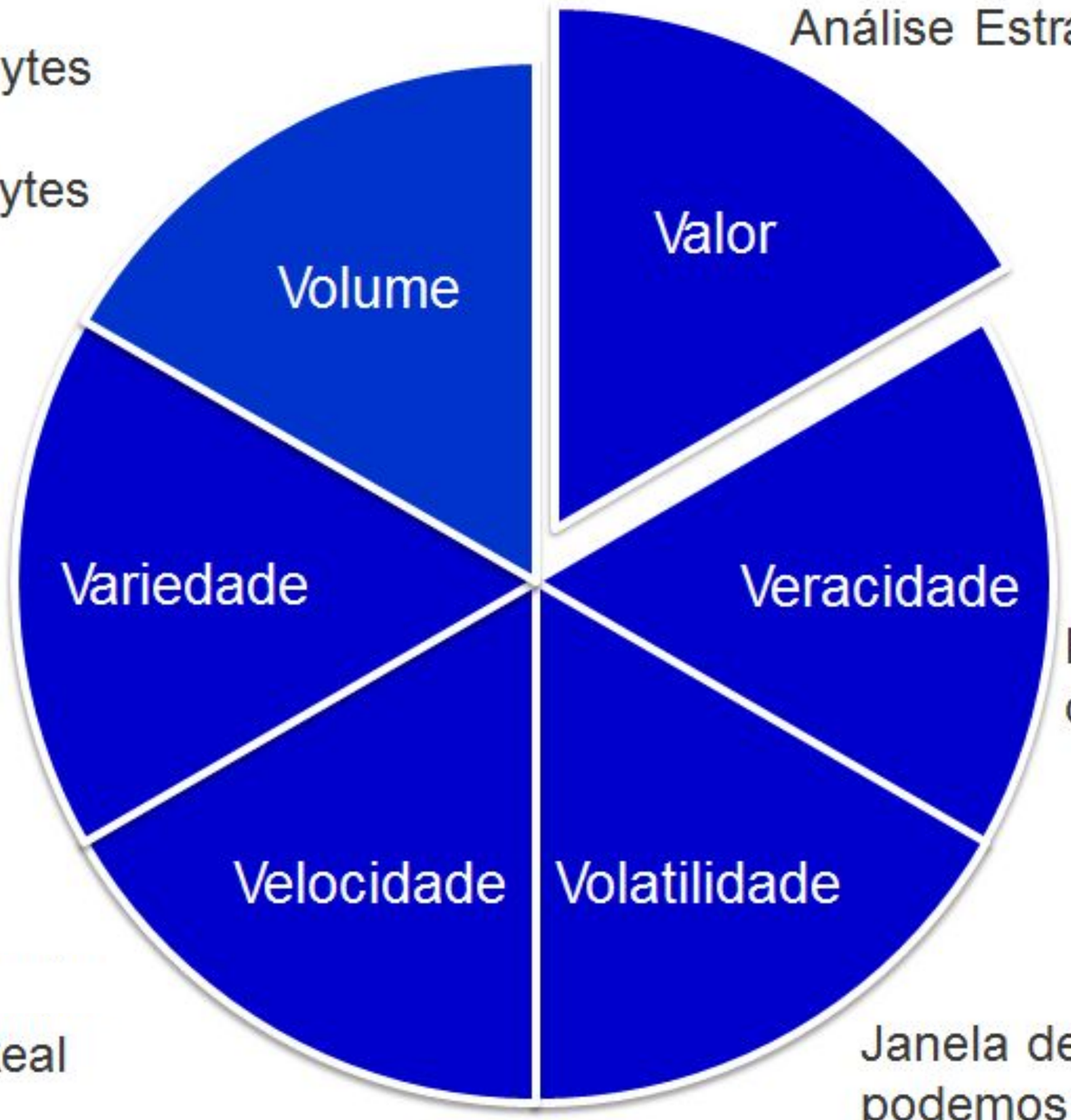
https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/

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

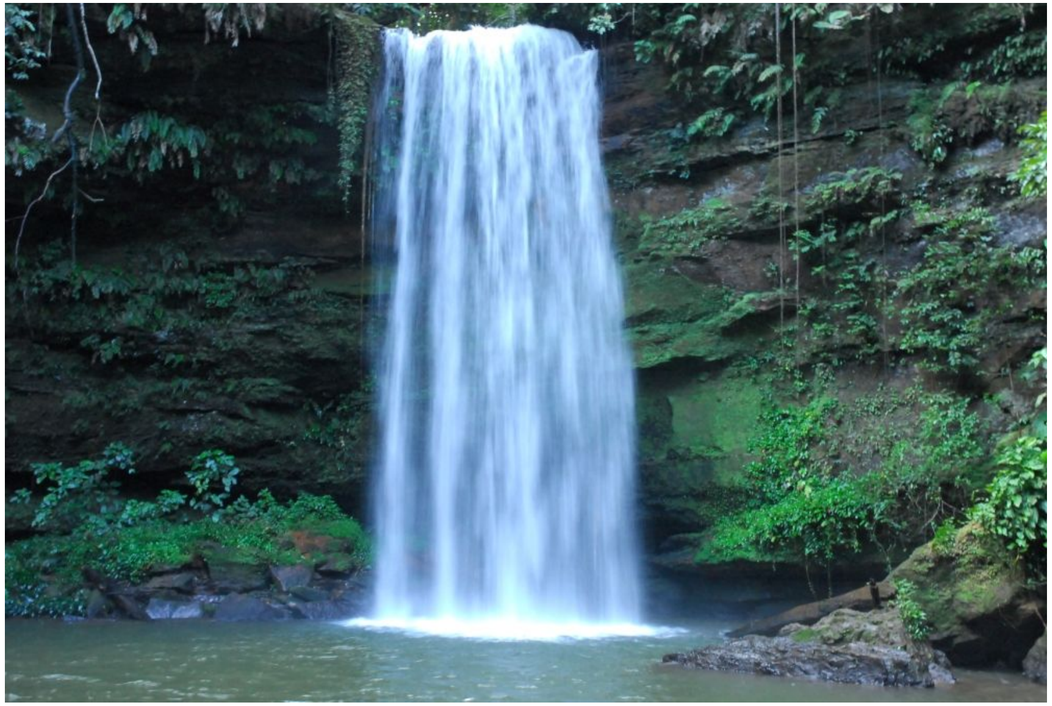
Batch
Tempo Real

Janela de tempo onde
podemos usar os dados

Streaming

Streaming

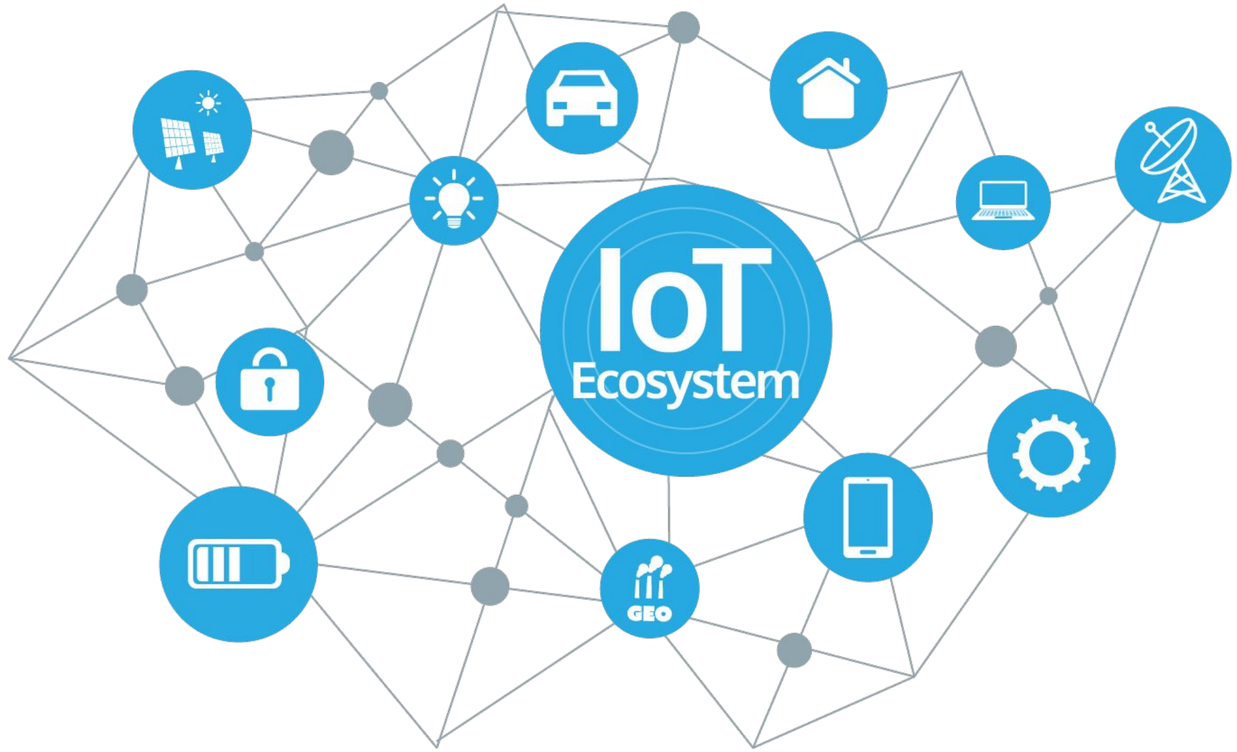
Fluxo contínuo (contínuo ≠ constante).



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 em Batch

Todas as entradas

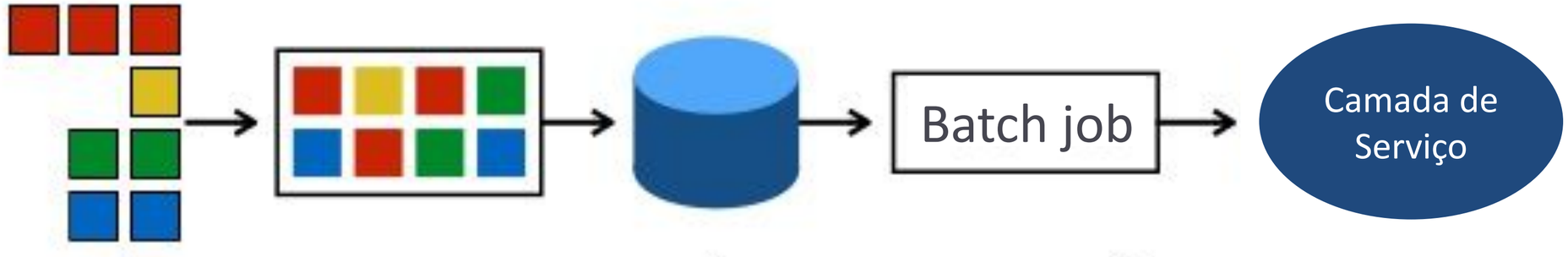
Hadoop,

Spark, etc.

Todas as saídas

Processamento de Streaming

Em geral:



continuamente produzido

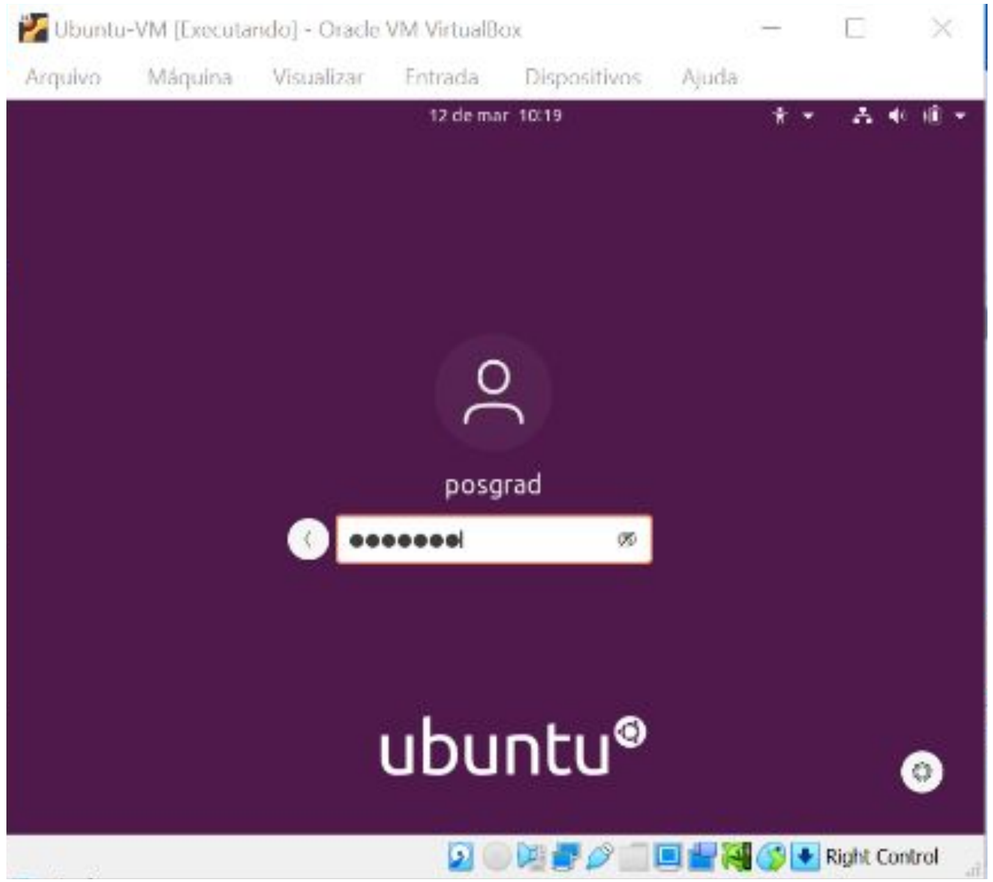
arquivos são **streams finitos**

periodicamente executado

Continuando nosso Setup...

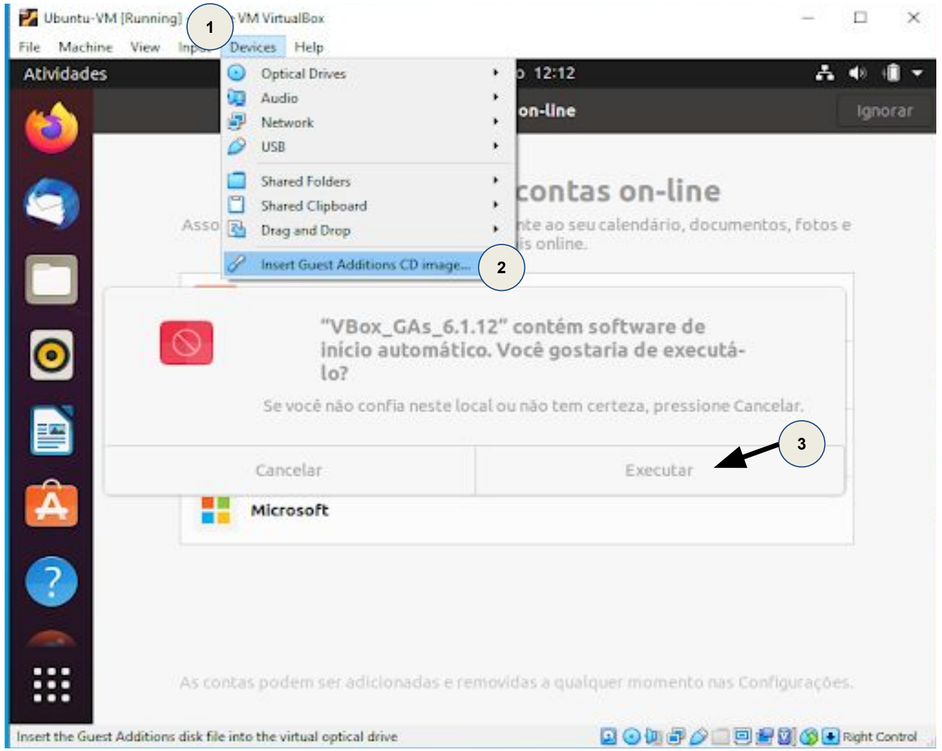
Configuração da VM Ubuntu

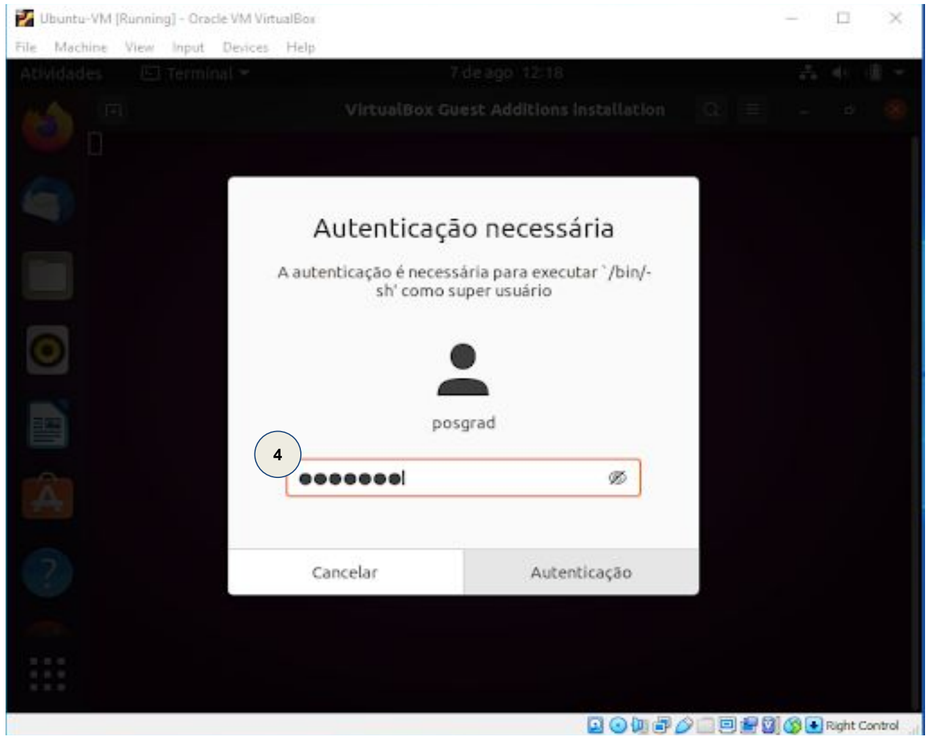
9. Logar na VM



Configuração da VM Ubuntu

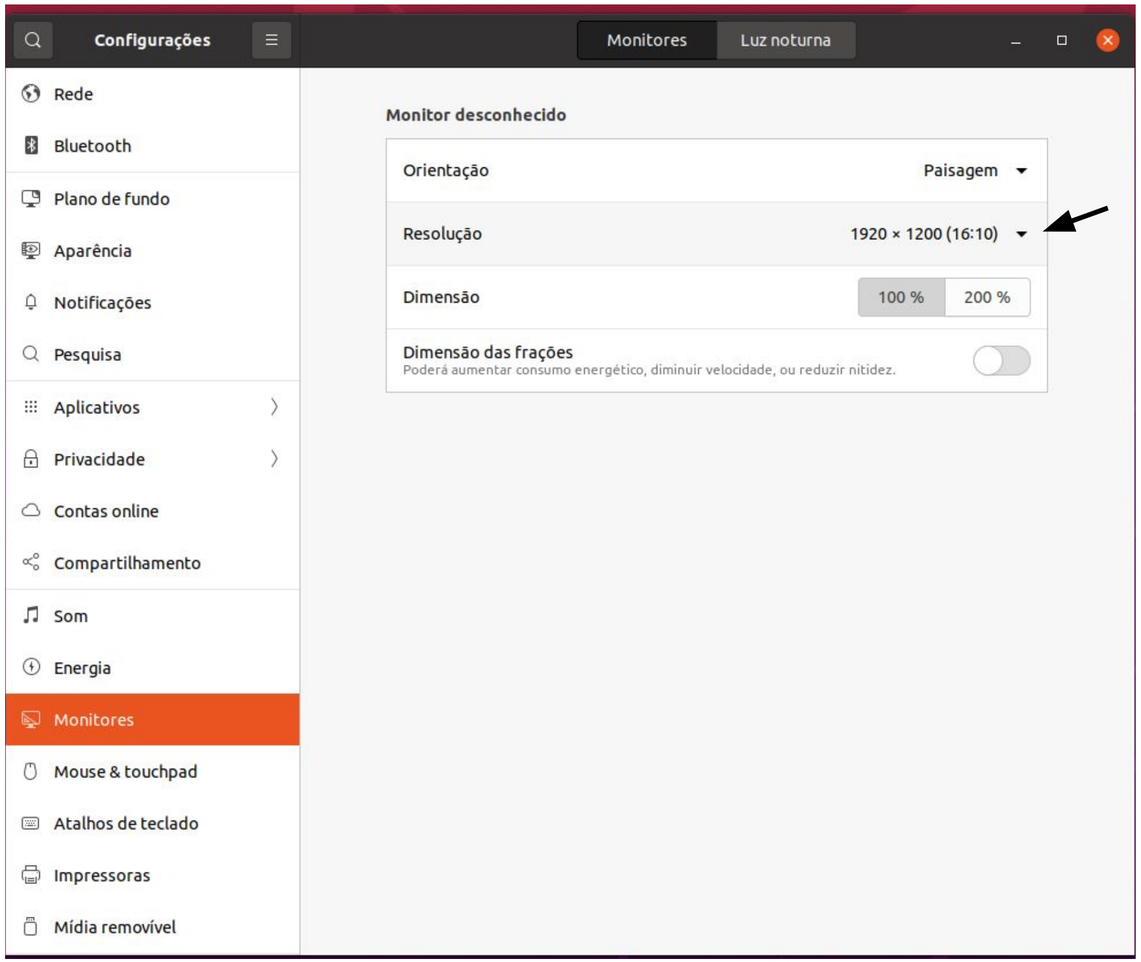
10. Instalar add-ons (para melhor experiência com o Ubuntu)





Configuração da VM Ubuntu

11. Ajustar resolução de tela



Atualizando o Ubuntu

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

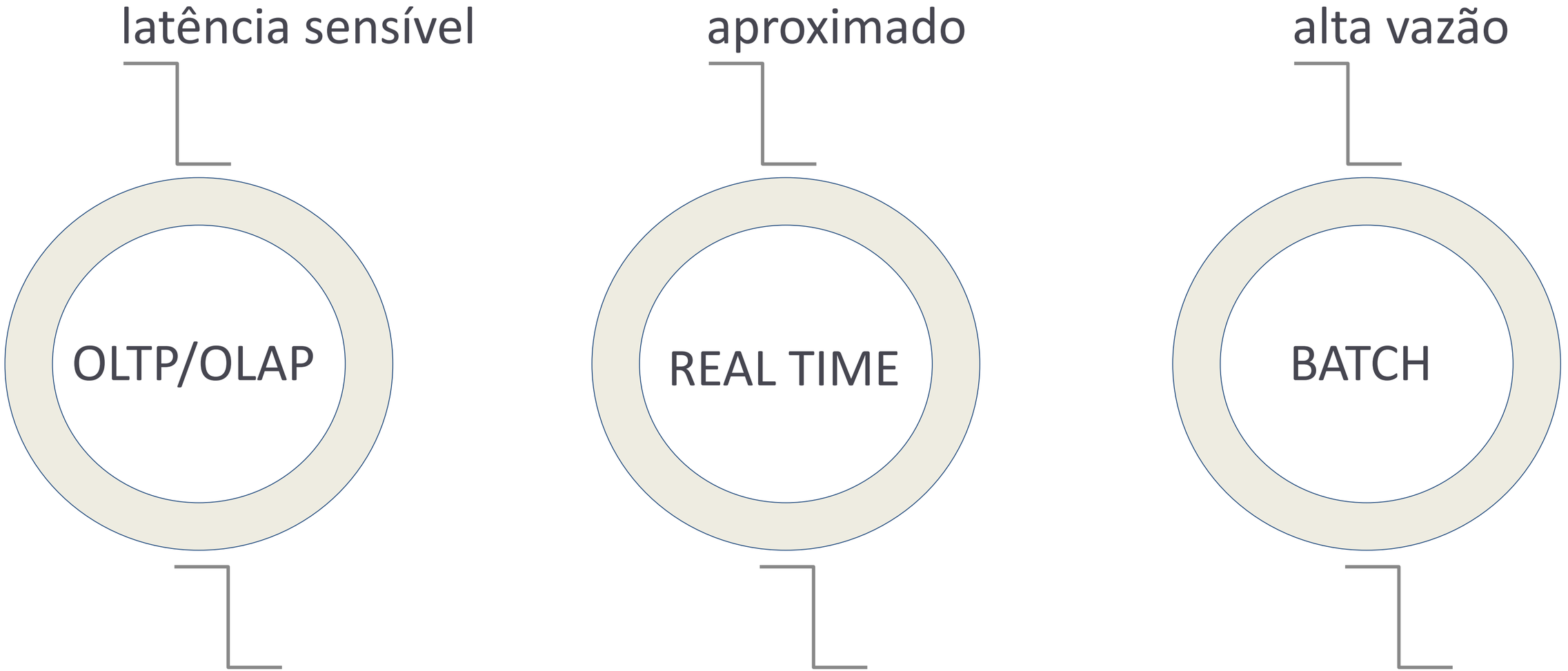
Atualizar o Ubuntu

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

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 nosso 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

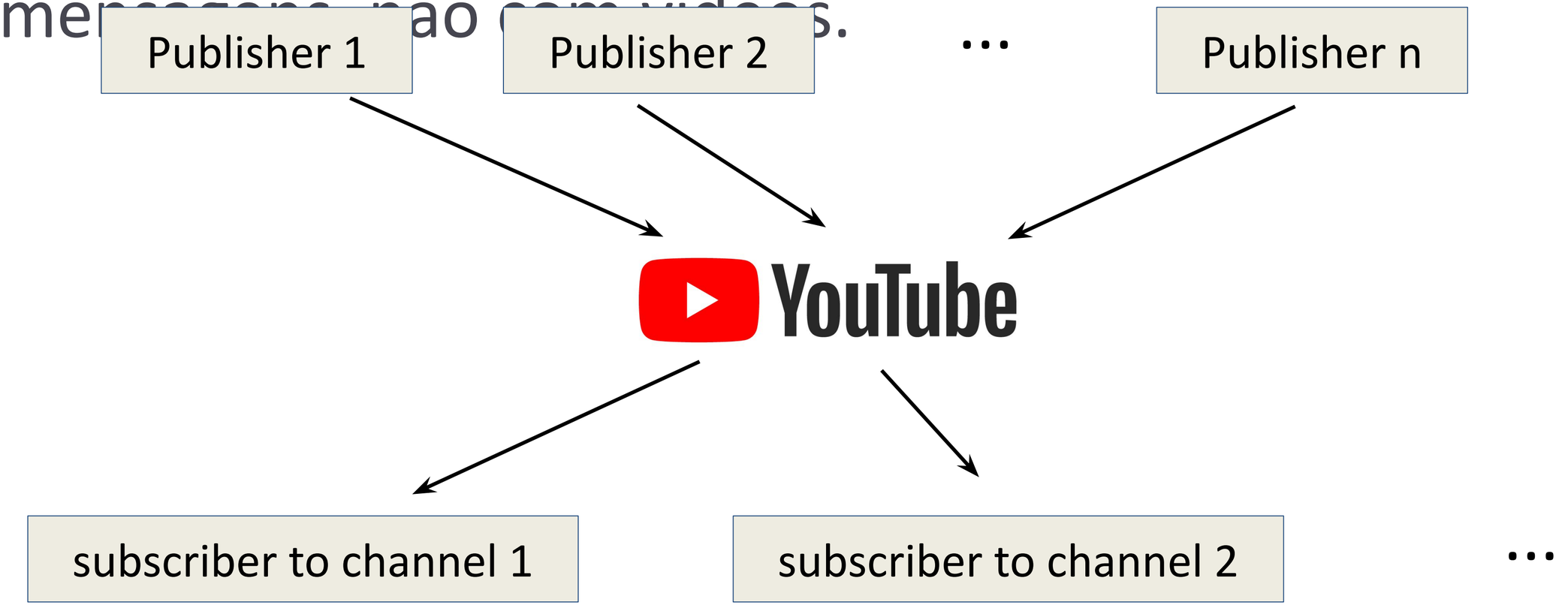
- Sistema de mensagens
  - Distribuído
  - Com alta vazão (*throughput*)
  - De geração (publicação) e leitura (sub-inscrição)
- Principais casos de uso:
  - Agregação de log
  - Mover/transformar conjuntos de dados em tempo real
  - Monitoramento

Apache Kafka

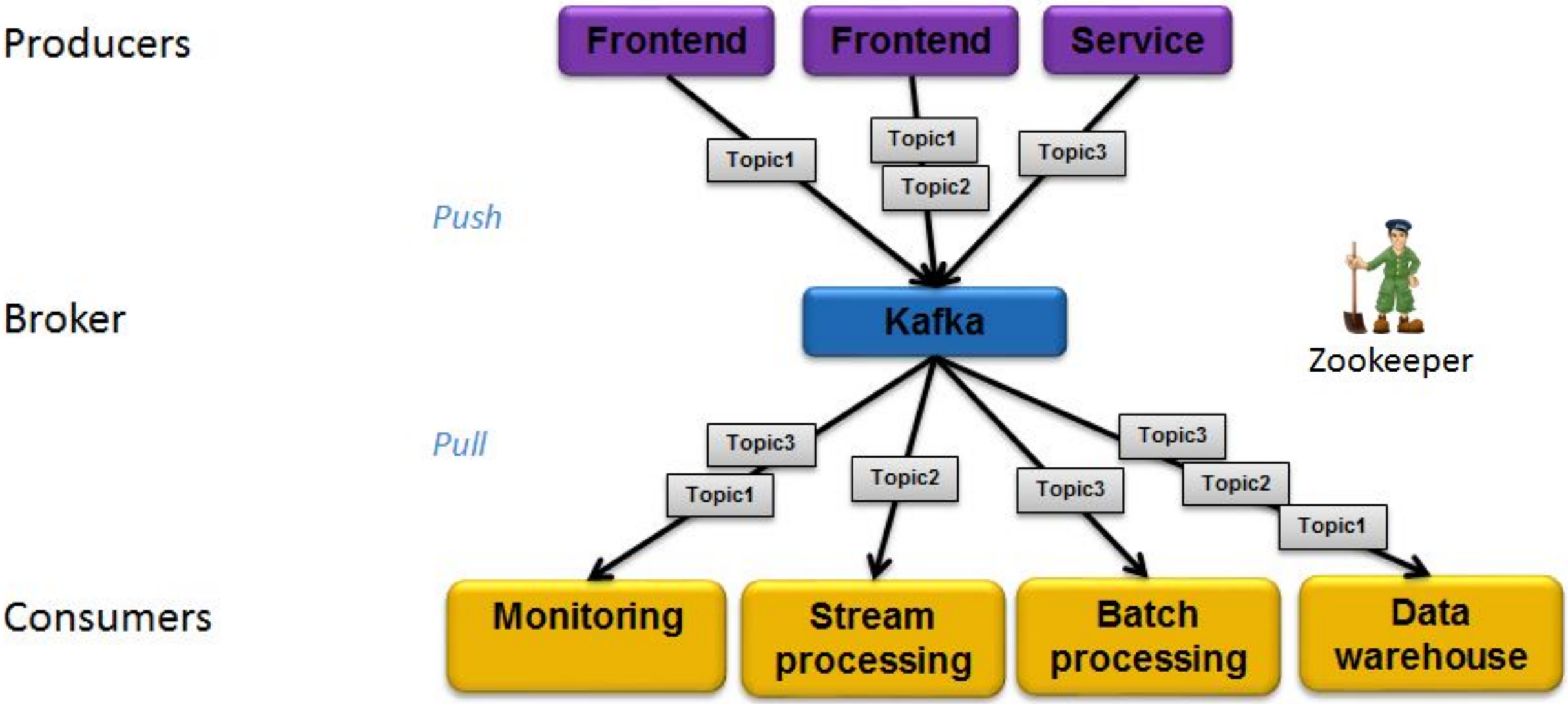
- Originalmente desenvolvido pelo LinkedIn.
- Implementado em scala/Java.
- *Producers* & *Consumers.*
- Mensagens são associadas a tópicos, os quais representam um stream específico.
  - Logs web
  - Dados de sensores
- *Consumers* se inscrevem em um ou mais tópicos.

Publisher-subscriber system

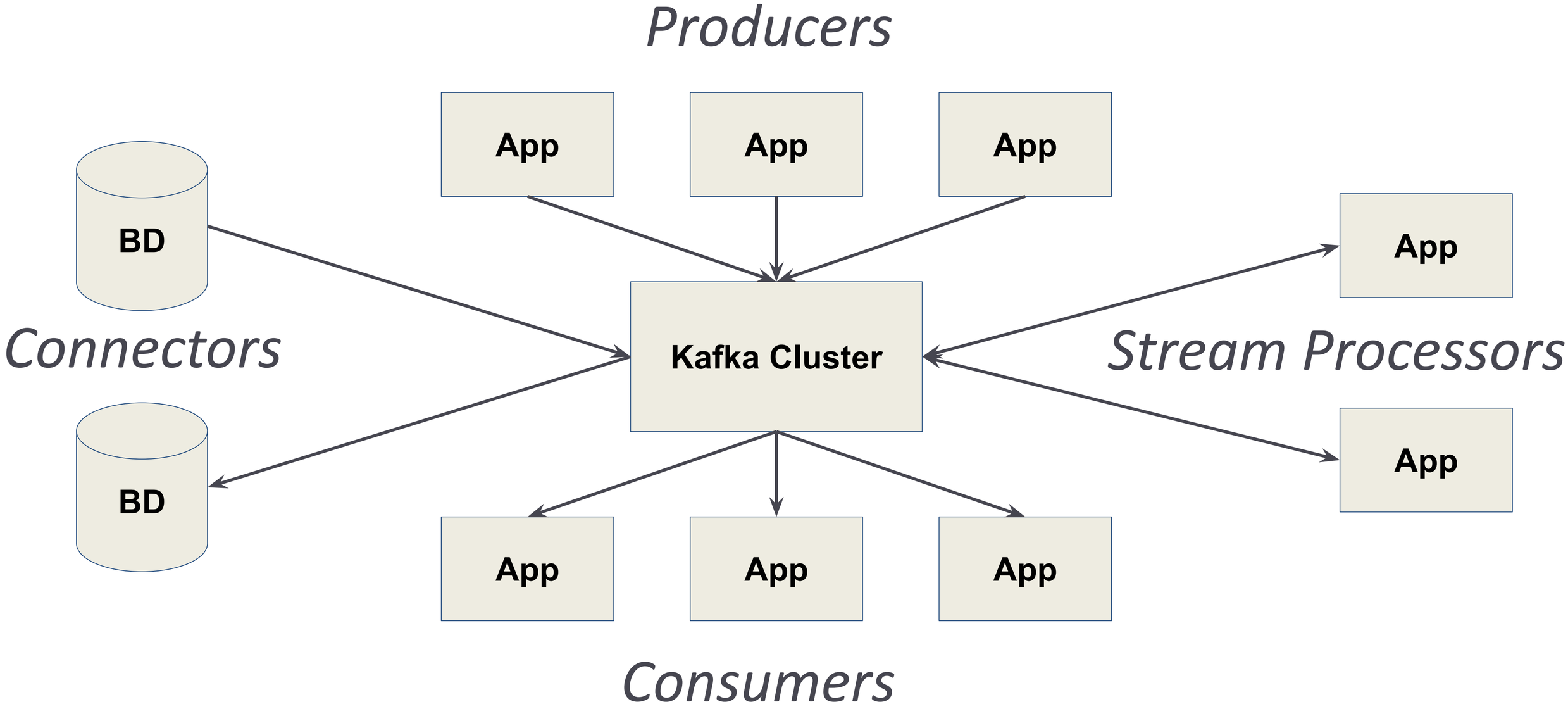
- Kafka pode ser visto como um sistema publisher/subscriber, como o Youtube, mas com mensagens não com vídeos.



Kafka: conceitos

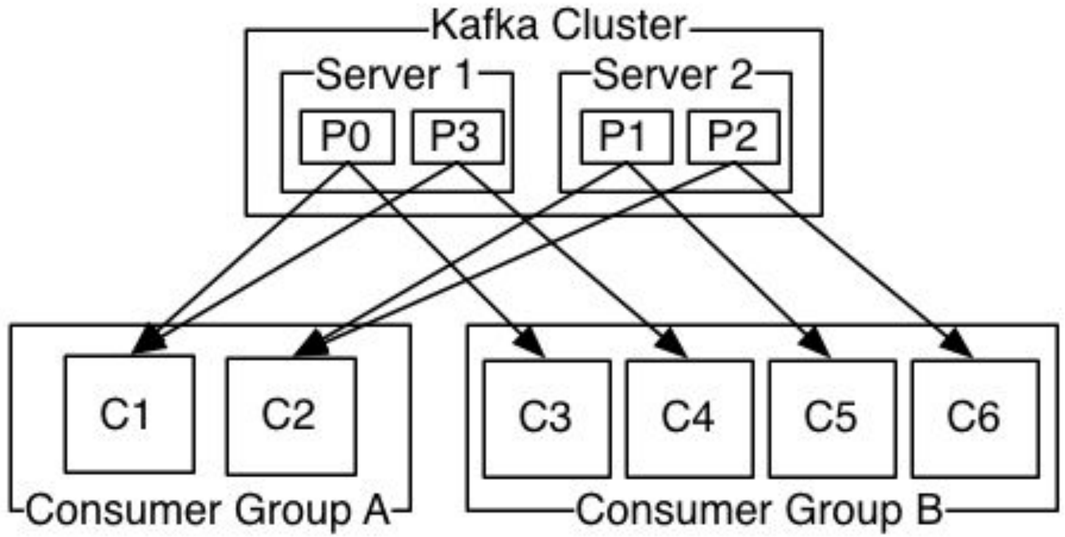


Kafka: arquitetura



Kafka: escalabilidade

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



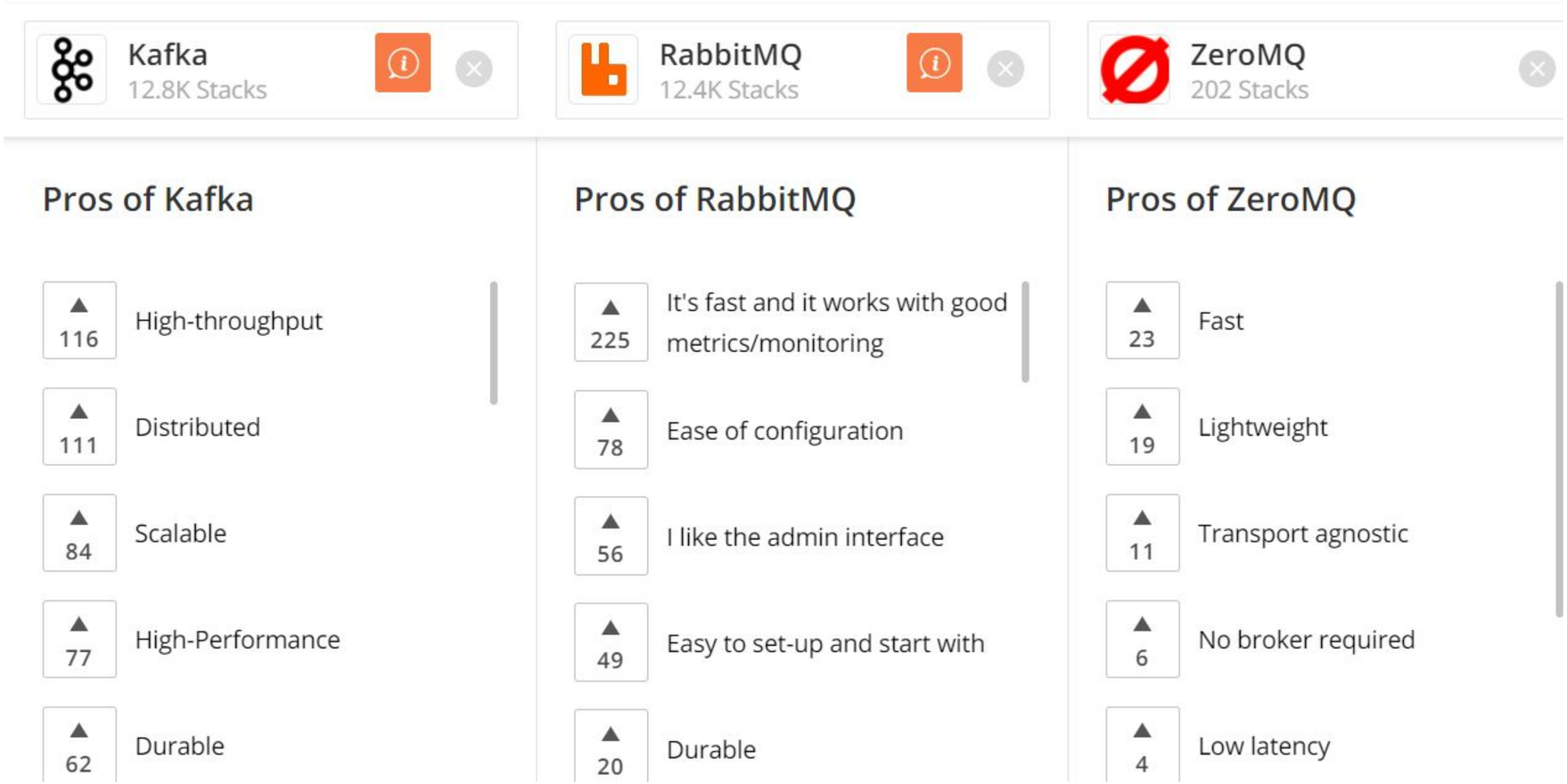
Fonte: https://kafka.apache.org/intro.html

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).

Kafka: quando comparado a outros concorrentes

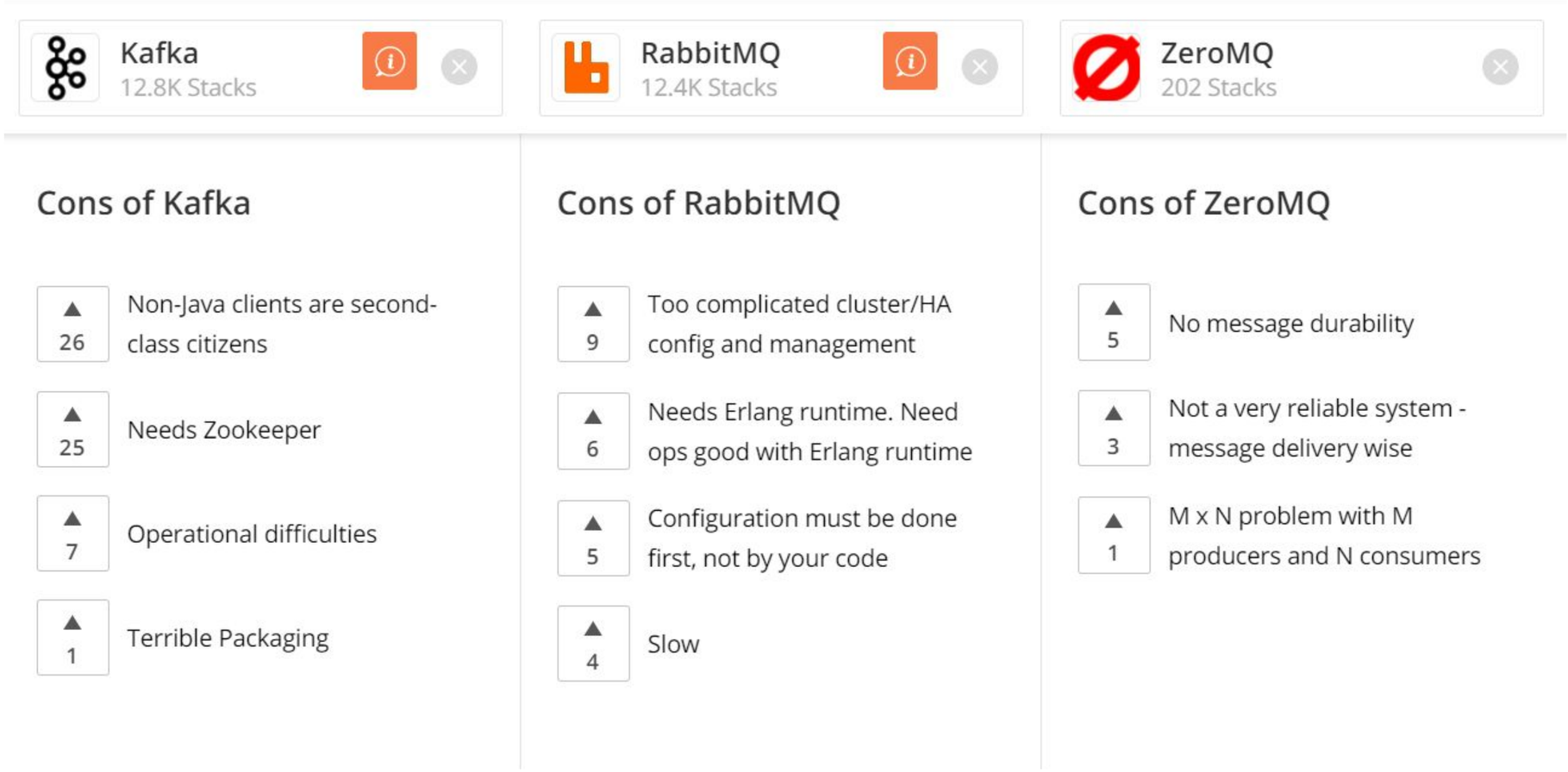
Prós



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

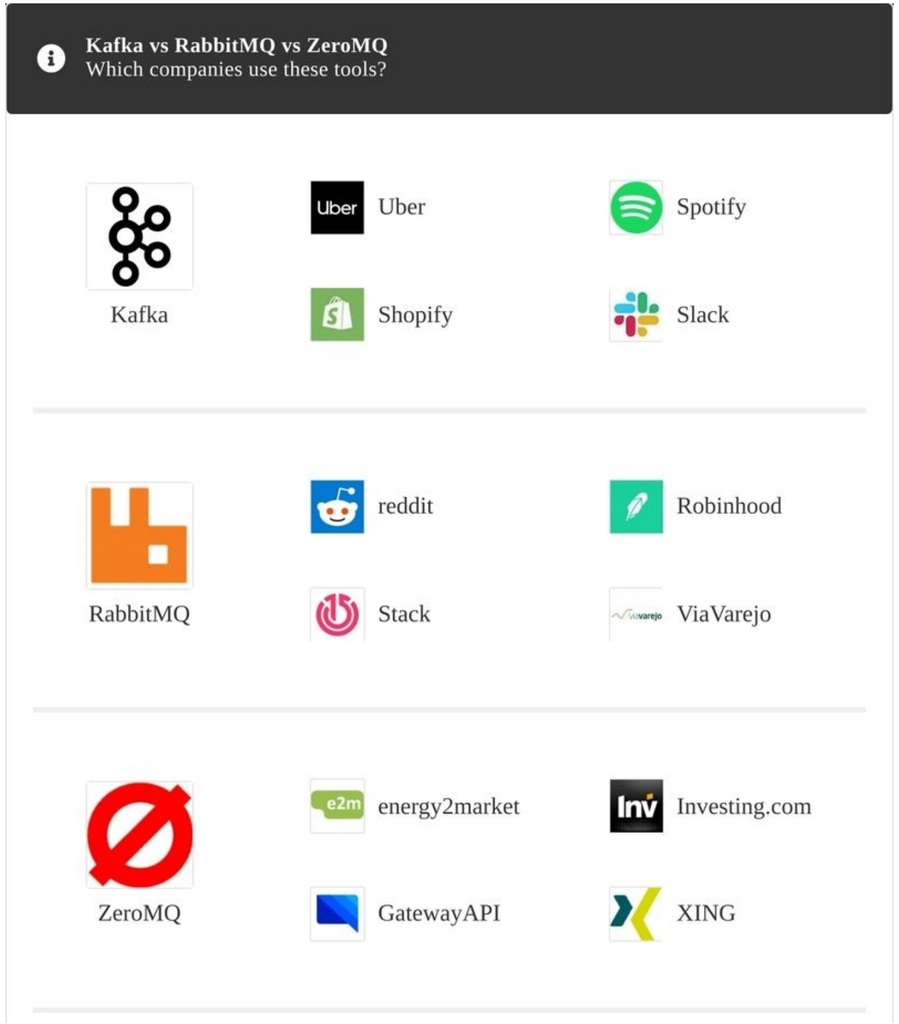
Kafka: quando comparado a outros concorrentes

Contras



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

Instalando o Apache Kafka

Realizar o download do Apache Kafka:

- `curl http://ftp.unicamp.br/pub/apache/kafka/2.8.1/kafka_2.12-2.8.1.tgz -o ~/Downloads/kafka.tgz`

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 62.6M  100 62.6M    0     0   528k      0  0:02:01  0:02:01 --:--:--  345k
posgrad@posgrad-vm:~$
```

Instalando o Apache Kafka

Criar um diretório chamado kafka e entrar nele:

- `mkdir kafka`
- `cd kafka`

Extrair os arquivos que estão na pasta download para o diretório criado:

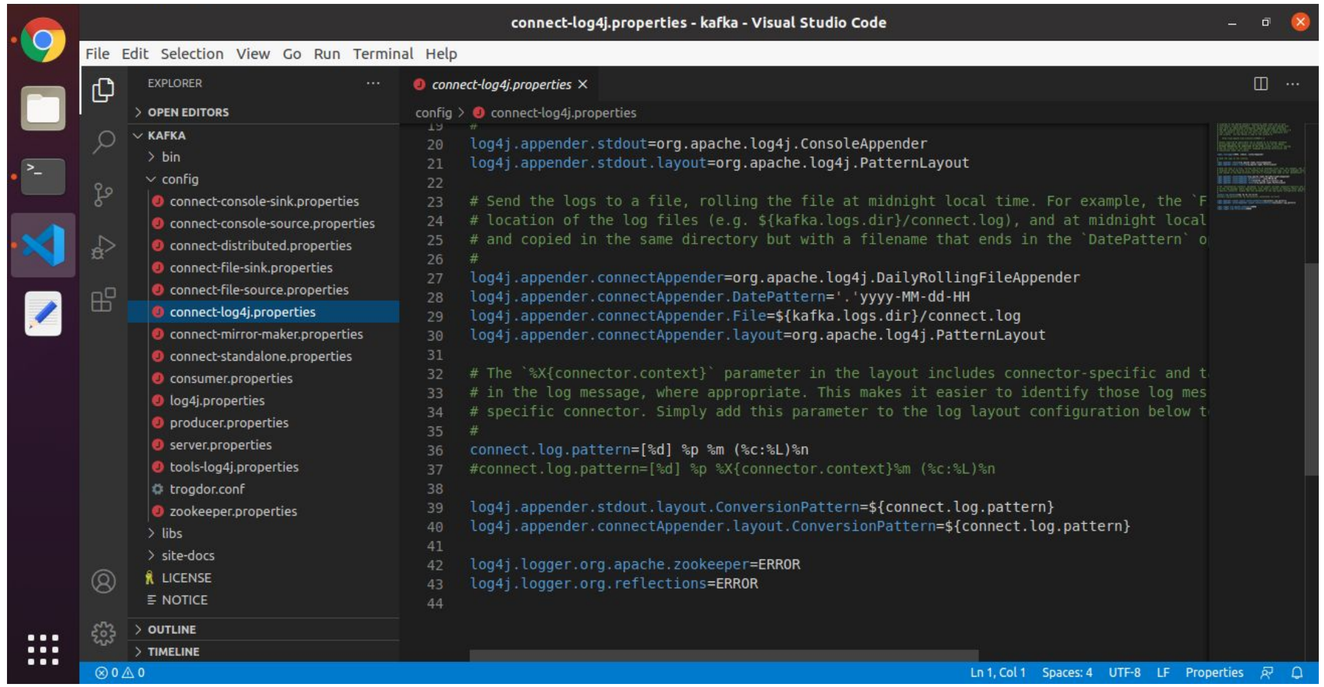
- `tar -xvzf ~/Downloads/kafka.tgz --strip 1`

Visualizando os arquivos do Kafka

Abrir o VS Code

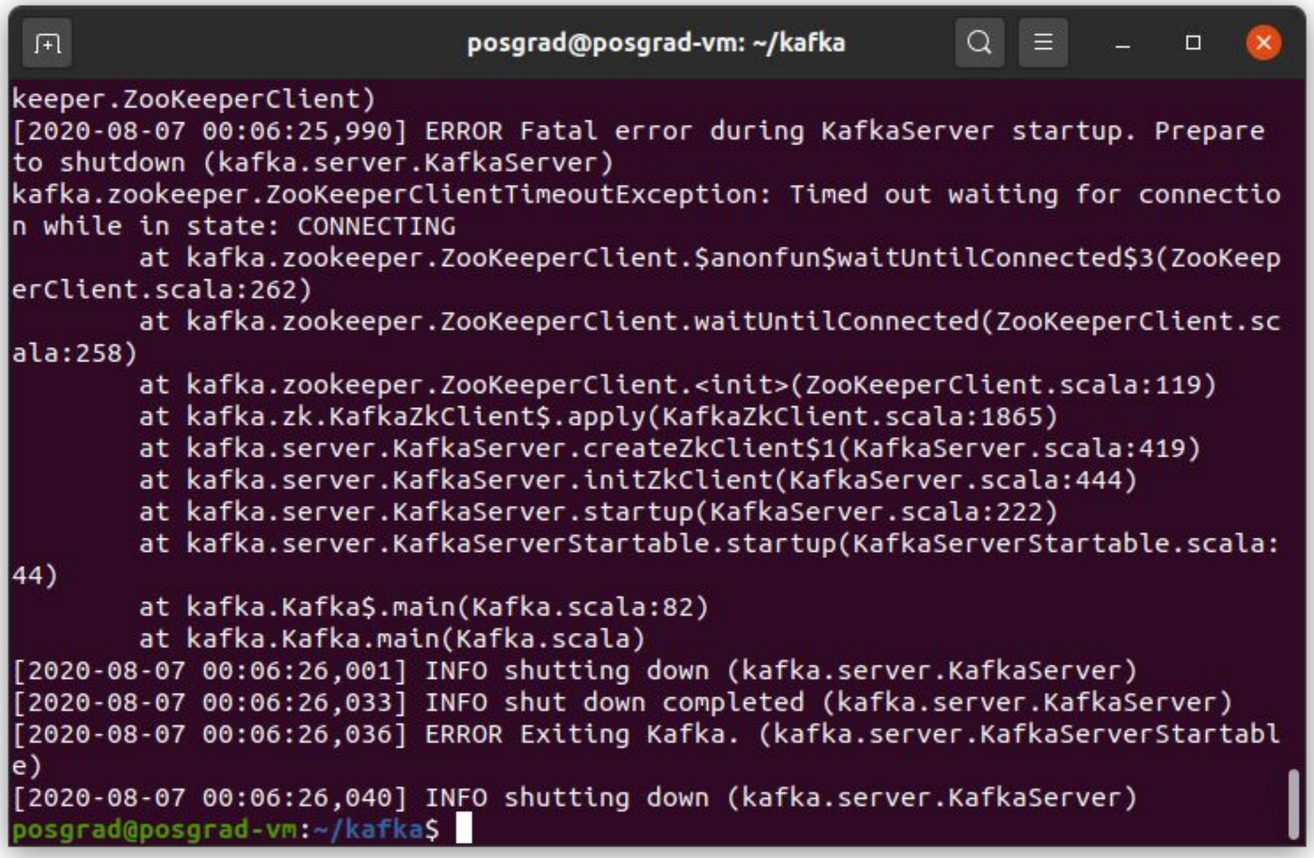
Clicar em Open Folder

Selecionar a pasta kafka



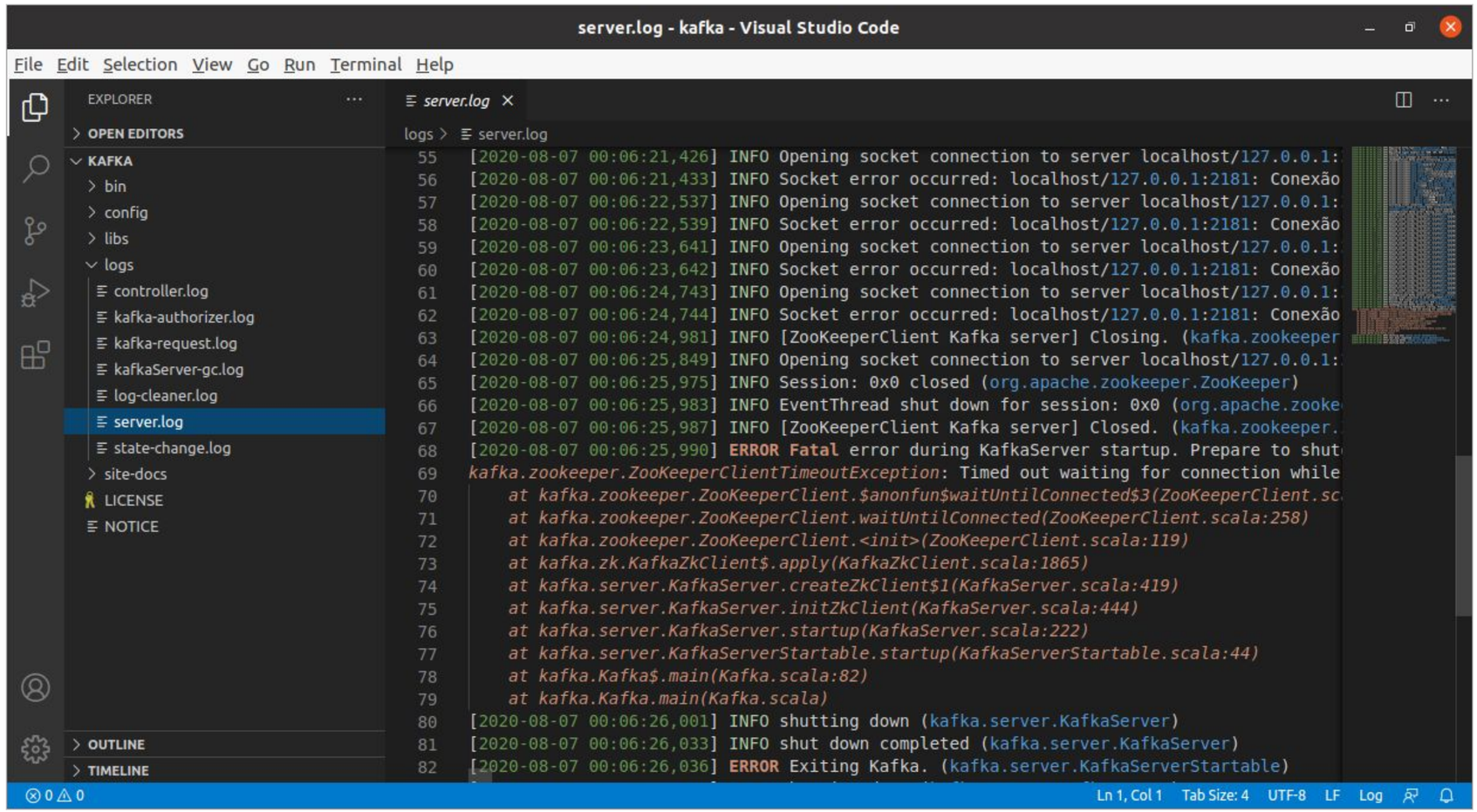
Inicializando o Servidor Kafka

➢ `bin/kafka-server-start.sh config/server.properties`



Observando os logs de erro do kafka

No VS Code, abrir a pasta logs, e o arquivo server.log

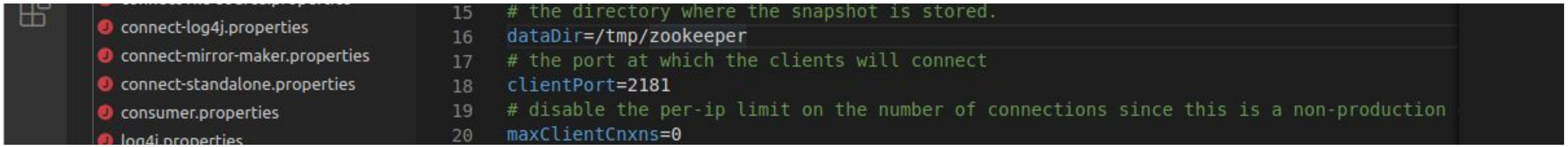


Inicializando o Zookeeper

```
➢ bin/zookeeper-server-start.sh config/zookeeper.properties
```

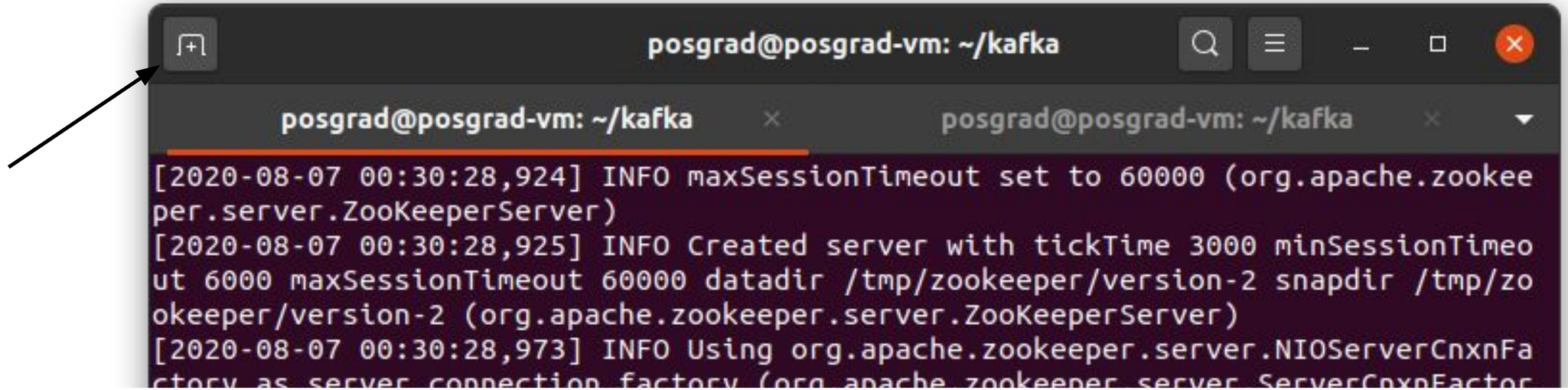
Zookeeper por padrão escuta a porta 2181

Observar arquivo ***zookeeper.properties*** na pasta ***config***



Inicializando o Servidor Kafka (novamente)

Em outro terminal



Inicie o Kafka novamente:

➢ `bin/kafka-server-start.sh config/server.properties`

Voilà

Zookeeper **localhost:2181**

Kafka server (broker) **localhost:9092**

Dúvidas?