

# Introdução às Ferramentas Python para Big Data

QXD0099 - Desenvolvimento de Software para Persistência

**Universidade Federal do Ceará - *Campus* Quixadá**

Prof. Francisco Victor da Silva Pinheiro  
victorpinheiro@ufc.br



# Agenda

- As 4 ondas
- O que é Big Data?
- Principais Aplicações
- Ferramentas Python para Big Data
  - Pandas – Manipulação de Dados
  - Dask – Processamento Paralelo
  - Apache Spark (PySpark) – Processamento Distribuído
  - SQLAlchemy + PostgreSQL – Armazenamento Estruturado
  - Apache Kafka – Processamento de Dados em Tempo Real
  - Comparação das Ferramentas
- Qual Ferramenta Escolher?

# As 4 ondas

- **Primeira Onda: Mobilidade (Mobile)**
  - Foco: Expansão dos dispositivos móveis e conectividade global
  - Impacto: Acesso a dados e serviços em qualquer lugar
  - Exemplos: Smartphones, aplicativos móveis, redes 4G/5G
- A primeira onda foi impulsionada pela revolução dos smartphones e tablets, tornando os dispositivos móveis a principal forma de acesso à internet. O crescimento do iOS e Android possibilitou a criação de um ecossistema de aplicativos que digitalizou setores inteiros.

# As 4 ondas

- **Segunda Onda: Social (Redes Sociais)**
  - Foco: Comunicação digital e redes sociais conectadas
  - Impacto: Mudança na forma como interagimos online
  - Exemplos: Facebook, Twitter, Instagram, TikTok
- A segunda onda foi marcada pelo surgimento e expansão das redes sociais, que mudaram completamente a forma como compartilhamos informações e interagimos.

# As 4 ondas

- **Terceira Onda: Cloud Computing (Computação em Nuvem)**
  - Foco: Processamento e armazenamento de dados online
  - Impacto: Redução de custos e escalabilidade para empresas
  - Exemplos: AWS, Google Cloud, Microsoft Azure
- A terceira onda veio com a computação em nuvem, permitindo que empresas armazenassem e processassem dados sem precisar de infraestrutura própria.

# As 4 ondas

- **Quarta Onda: Big Data e Inteligência Artificial**
  - Foco: Extração de valor dos dados para decisões inteligentes
  - Impacto: Automação, aprendizado de máquina e previsões avançadas
  - Exemplos: Machine Learning, IA Generativa, Análise de Dados
- A quarta onda é impulsionada pela explosão de dados e o uso de Inteligência Artificial para extrair valor dessas informações.



# O que é Big Data?

- **Volume:** Quantidade massiva de dados.
- **Velocidade:** Geração contínua e rápida.
- **Variedade:** Dados estruturados (bancos de dados), semiestruturados (JSON, XML) e não estruturados (vídeos, imagens).
- **Veracidade:** Qualidade e confiabilidade dos dados.





# Principais Aplicações

- Processamento de logs em tempo real (observabilidade de sistemas).
- Análise de grandes volumes de dados financeiros.
- Treinamento de modelos de Machine Learning.
- Extração de insights a partir de redes sociais.

# Setores com maior potencial de benefício

- Saúde
- Governo
- Comércio
- Indústria
- Tecnologia de localização pessoal

# Ferramentas Python para Big Data

- Python é amplamente utilizado em Big Data devido à sua facilidade de uso, diversidade de bibliotecas e integração com diversas plataformas de armazenamento e processamento distribuído.



# Pandas – Manipulação de Dados

- Pandas é uma biblioteca essencial para manipulação e análise de dados. Permite trabalhar com tabelas de dados (DataFrames) de forma eficiente.
- Principais Recursos:
  - Leitura de arquivos grandes (.csv, .json, .parquet).
  - Manipulação de colunas e indexação.
  - Agregações e estatísticas descritivas.



# Pandas – Exemplo de uso

```
import pandas as pd

# Leitura de um grande arquivo CSV
df = pd.read_csv("dados.csv", chunksize=10000) # Leitura em blocos de 10.000 linhas

# Processamento de cada bloco
for chunk in df:
    print(chunk.describe()) # Estatísticas básicas
```

# Dask – Processamento Paralelo

- Dask é uma biblioteca que permite o processamento de dados maiores que a memória, distribuindo a carga de trabalho.
- Por que usar Dask?
  - Processa grandes DataFrames de forma paralela.
  - Utiliza o mesmo formato de Pandas, facilitando a transição.



# Dask – Exemplo de uso

```
import dask.dataframe as dd

# Leitura de um arquivo grande usando Dask
df = dd.read_csv("dados_grandes.csv")

# Operação em paralelo
df_mean = df.groupby("categoria")["valor"].mean().compute()
print(df_mean)
```

# Apache Spark (PySpark) – Processamento Distribuído

- O Apache Spark é uma das principais ferramentas para Big Data. Com o PySpark, podemos utilizar o Spark diretamente no Python.
- Benefícios do Spark:
  - Processamento distribuído eficiente.
  - Suporte a SQL, Machine Learning e Streaming.
  - Compatível com Hadoop e armazenamento na nuvem.





# Apache Spark (PySpark) – Exemplo de uso

```
from pyspark.sql import SparkSession

# Criar uma sessão Spark
spark = SparkSession.builder.appName("BigData").getOrCreate()

# Leitura de um grande CSV em formato Spark DataFrame
df = spark.read.csv("dados.csv", header=True, inferSchema=True)

# Contagem de registros
print(df.count())
```

# SQLAlchemy + PostgreSQL – Armazenamento Estruturado

- Para Big Data, bancos de dados relacionais como PostgreSQL são frequentemente utilizados, especialmente com suporte a armazenamento colunar e otimização para grandes volumes de dados.



# Apache Kafka – Processamento de Dados em Tempo Real

- O Apache Kafka é uma plataforma de streaming distribuída e de código aberto usada para processar e gerenciar fluxos de dados em tempo real. Originalmente desenvolvido pelo LinkedIn, foi doado à Apache Software Foundation, onde se tornou um dos sistemas mais populares para manipulação de grandes volumes de dados.
- Kafka é amplamente utilizado para ingestão, processamento e distribuição de dados em tempo real e é uma peça fundamental em arquiteturas de Big Data e IoT.



# Apache Kafka – Exemplo de uso

```
from kafka import KafkaProducer
import json

producer = KafkaProducer(bootstrap_servers='localhost:9092',
                          value_serializer=lambda v: json.dumps(v).encode('utf-8'))

# Enviando uma mensagem
producer.send("topico_dados", {"id": 1, "valor": 100})
producer.flush()
```

# Comparação das Ferramentas

Ferramenta	Principal Uso	Tipo de Processamento
Pandas	Análise de dados	Memória RAM (Single Machine)
Dask	Processamento paralelo	Paralelo (Multi-core)
PySpark	Processamento distribuído	Cluster de Máquinas
PostgreSQL + SQLAlchemy	Banco de dados estruturado	Armazenamento + Queries
Kafka	Processamento de eventos	Streaming de Dados

# Qual Ferramenta Escolher?

- A escolha da ferramenta depende do cenário:
  - Pequenos volumes de dados → Pandas.
  - Grande volume de dados em única máquina → Dask.
  - Cluster e processamento distribuído → PySpark.
  - Armazenamento estruturado e consultas SQL → PostgreSQL + SQLAlchemy.
  - Streaming de dados e eventos em tempo real → Kafka

# Referências

- Você realmente sabe o que é Big Data?
  - [https://www.ibm.com/developerworks/mydeveloperworks/blog/s/ctaurion/entry/voce\\_realmente\\_sabe\\_o\\_que\\_e\\_big\\_data?g=en](https://www.ibm.com/developerworks/mydeveloperworks/blog/s/ctaurion/entry/voce_realmente_sabe_o_que_e_big_data?g=en)
- A quantas anda o Big Data no atual mercado de tecnologia?
  - <http://imasters.com.br/banco-de-dados/a-quantas-anda-o-big-data-no-atual-mercado-de-tecnologia/>
- Big data é um tsunami em alto mar – Resenha do livro Big Data
  - <http://imasters.com.br/tecnologia/redes-e-servidores/resenha-do-livro-big-data/>



# Obrigado!

## Dúvidas?



**Universidade Federal do Ceará - *Campus* Quixadá**

Prof. Francisco Victor da Silva Pinheiro  
victorpinheiro@ufc.br

