Introdução ao PySpark

Preparando os Dados para Atender ao Negócio

Muito Prazer

Juan Eduardo Domingos

Formação

Bacharel em Ciência da Computação pelo Centro Universitário de Jaguariúna - UniFAJ

Pós Graduado em Sistemas Distribuídos pela Pontifícia Universidade Católica do Paraná - PUCPR

Carreira

9 anos trabalhando com tecnologia

6 anos na área de dados

Atualmente Lead Data Engineer na WinDifferent uma companhia americana de Growth e Prospects

https://www.linkedin.com/in/juan-domingos/

https://github.com/jeduardodomingos

Agenda

- **Introdução**
  - Introdução ao Apache Spark
  - Arquitetura Base
- **Manipulando Dados com PySpark**
  - Leitura de Dados
  - SQL Operations

O que é o Apache Spark

- Processamento Distribuído

- Processamento em Memória

- Até 100 vezes Mais Rápido que o Hadoop Map-Reduce

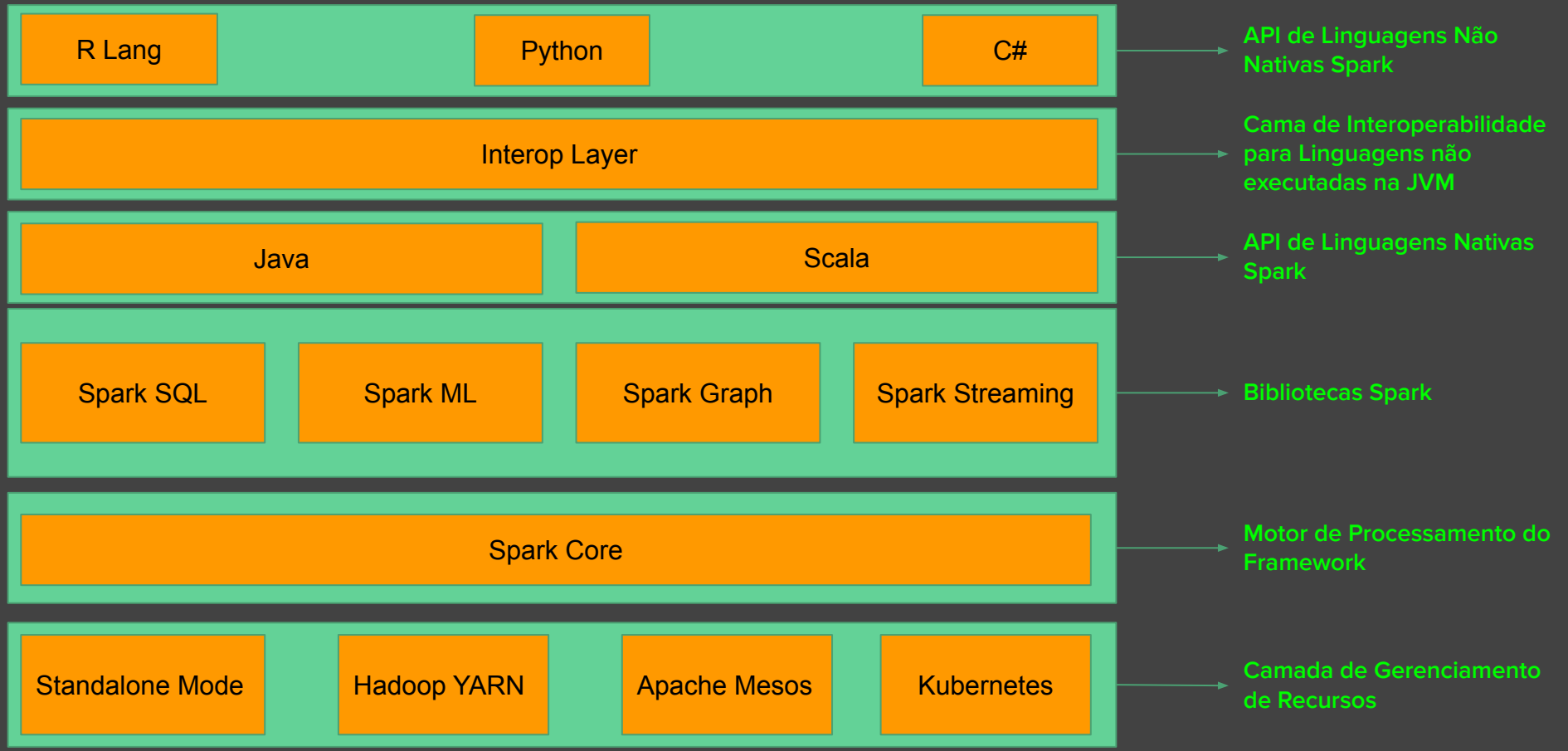
O que é o Apache Spark

- Pode ser Executado Sobre um Cluster ou em uma única Máquina

- Possui Interfaces que permitem trabalhar com: Grafos, Machine Learning, SQL, além de estruturas Streaming

- Altamente aderente a ambientes Cloud

Arquitetura Base



Vamos Botar a Mão na Massa

Dúvidas?

Obrigado