

Processamento de Big Data com PySpark

Trabalhando com Dados em Larga Escala

Lis Barreto



Lis Barreto

Engenheira de Machine Learning no Itaú Unibanco

- Especialista em Machine Learning Engineering
- Especialista em DevOps & Software Engineering
- Graduada em Banco de Dados
- Membro da Comunidade Python
- Ex-UFS (Física e Engenharia de Computação)



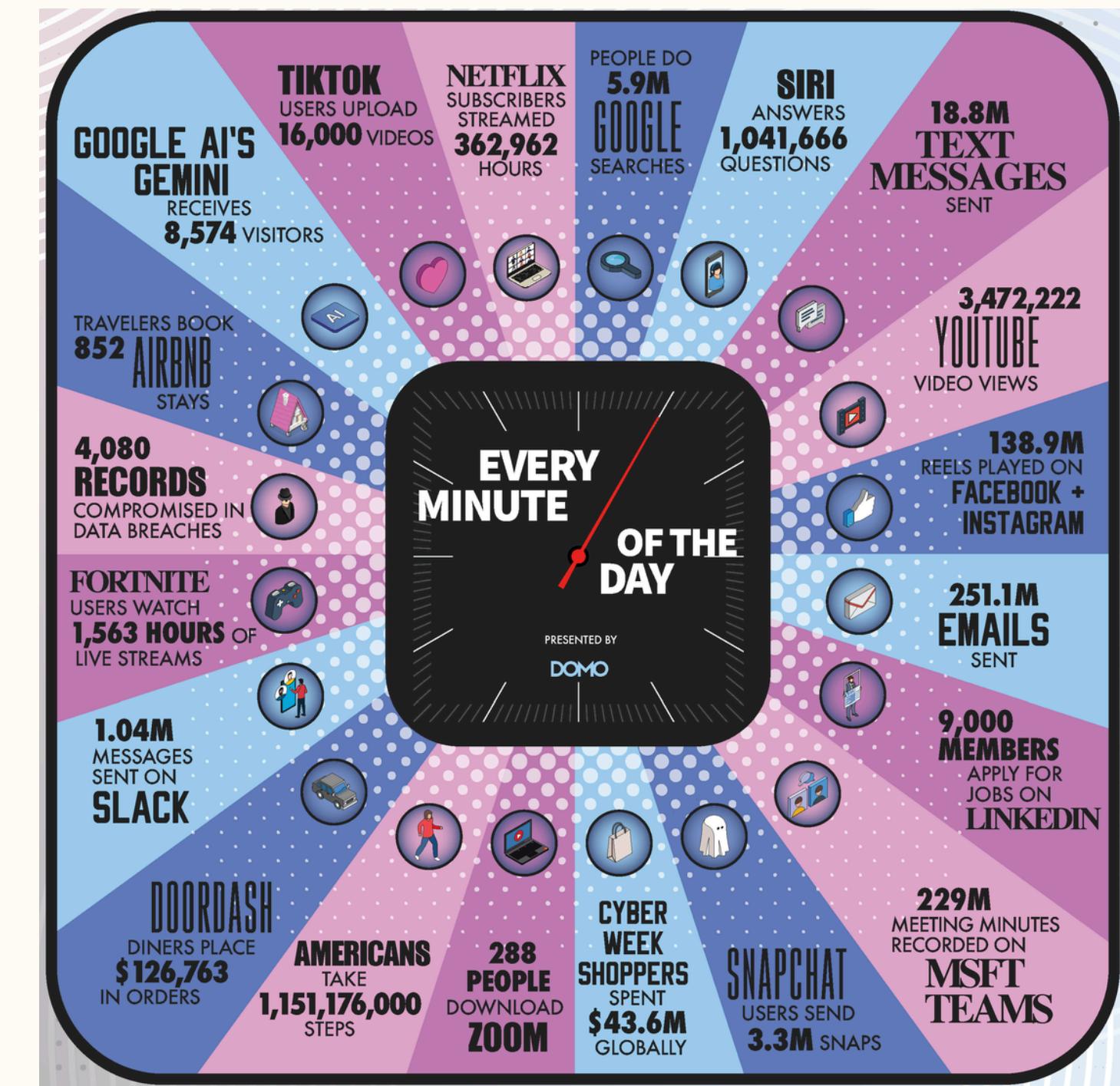
O Universo dos Dados

Volume, Ferramentas e Necessidade Estratégica

Os Dados Nunca Dormem

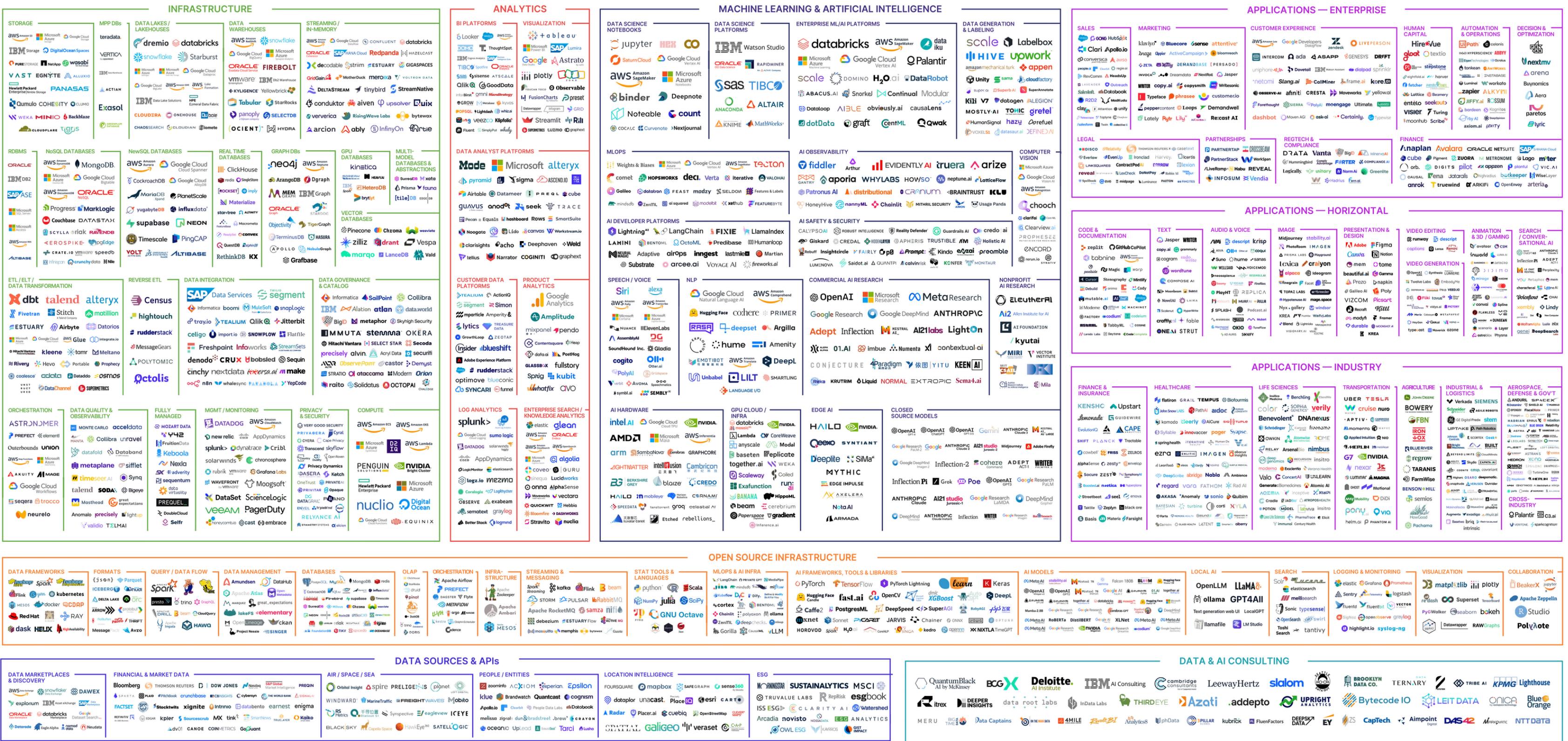
Em 2024, 5,52 bilhões de pessoas (67,5% da população global) estiveram online.

O volume total de dados criados, capturados e consumidos globalmente deve atingir 394 zettabytes até 2028.



Fonte: Infográfico "Data Never Sleeps 12.0". Criado anualmente pela Domo.

THE 2024 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE



Version 1.0 - March 2024

© Matt Turck (@mattturck), Aman Kabeer (@AmanKabeer11) & FirstMark (@firstmarkcap)

Blog post: mattturck.com/MAD2024

Interactive version: MAD.firstmarkcap.com

Comments? Email MAD2024@firstmarkcap.com

FIRST MARK
EARLY STAGE VENTURE CAPITAL

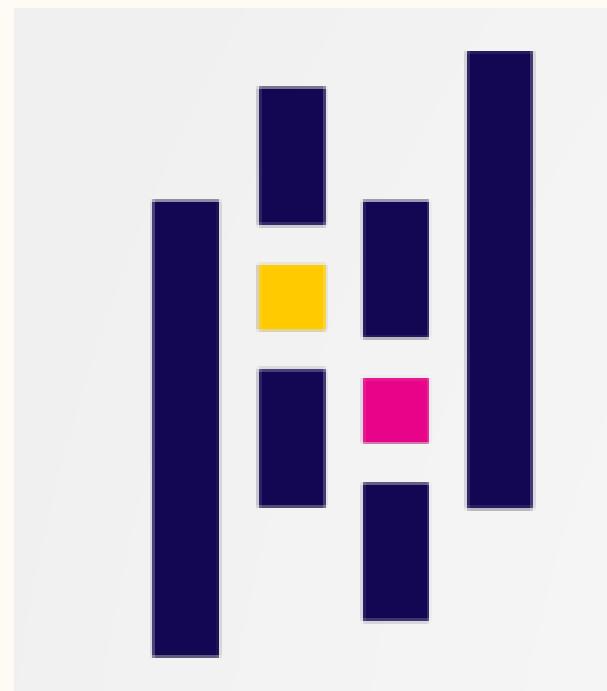
Fonte: Landscape do ecossistema de Dados, Analytics, Machine Learning e IA. Criado anualmente por Matt Turck.



Processamento Paralelo e Distribuído

Como o Spark utiliza o conceito de partições para um processamento de dados mais eficiente

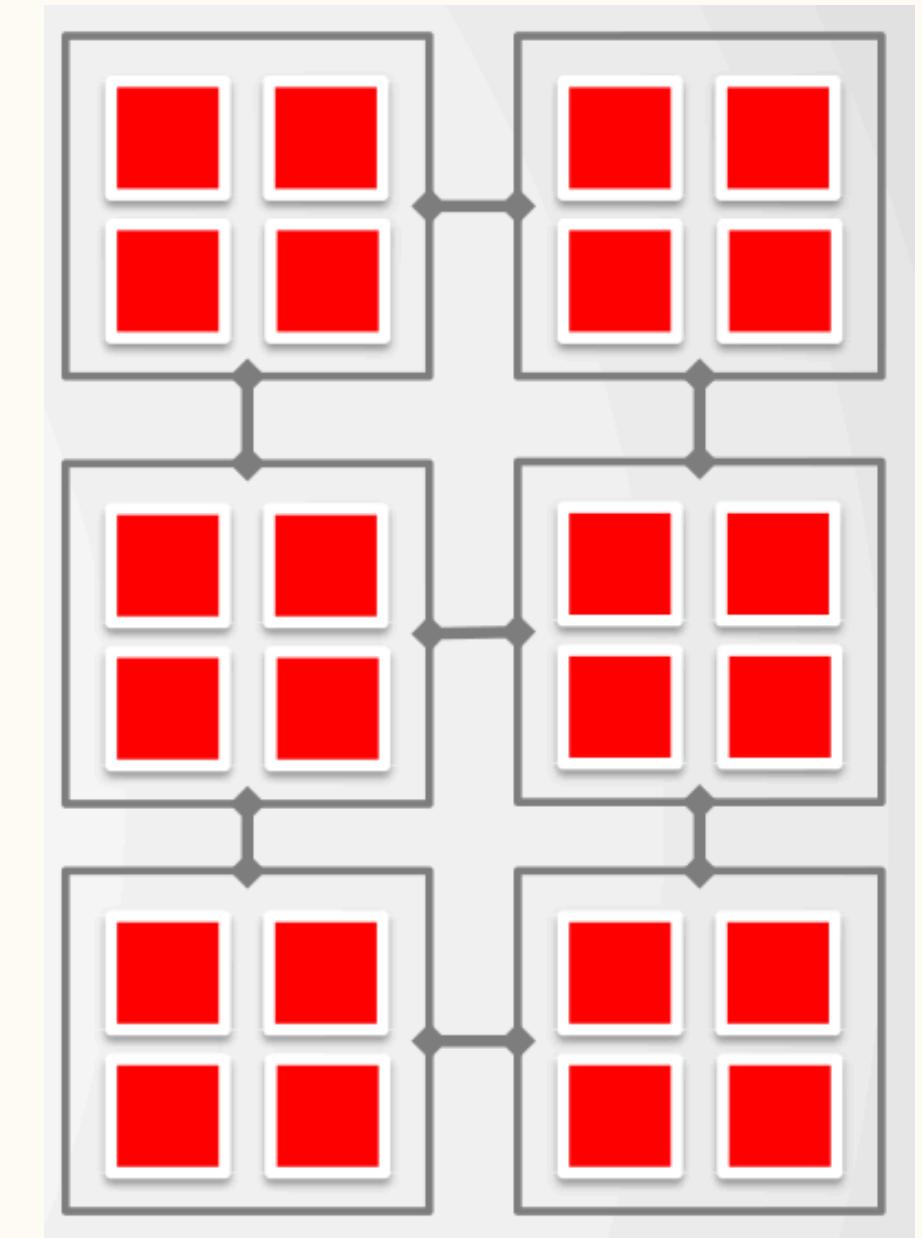
Pandas vs PySpark



Pandas: Single Node

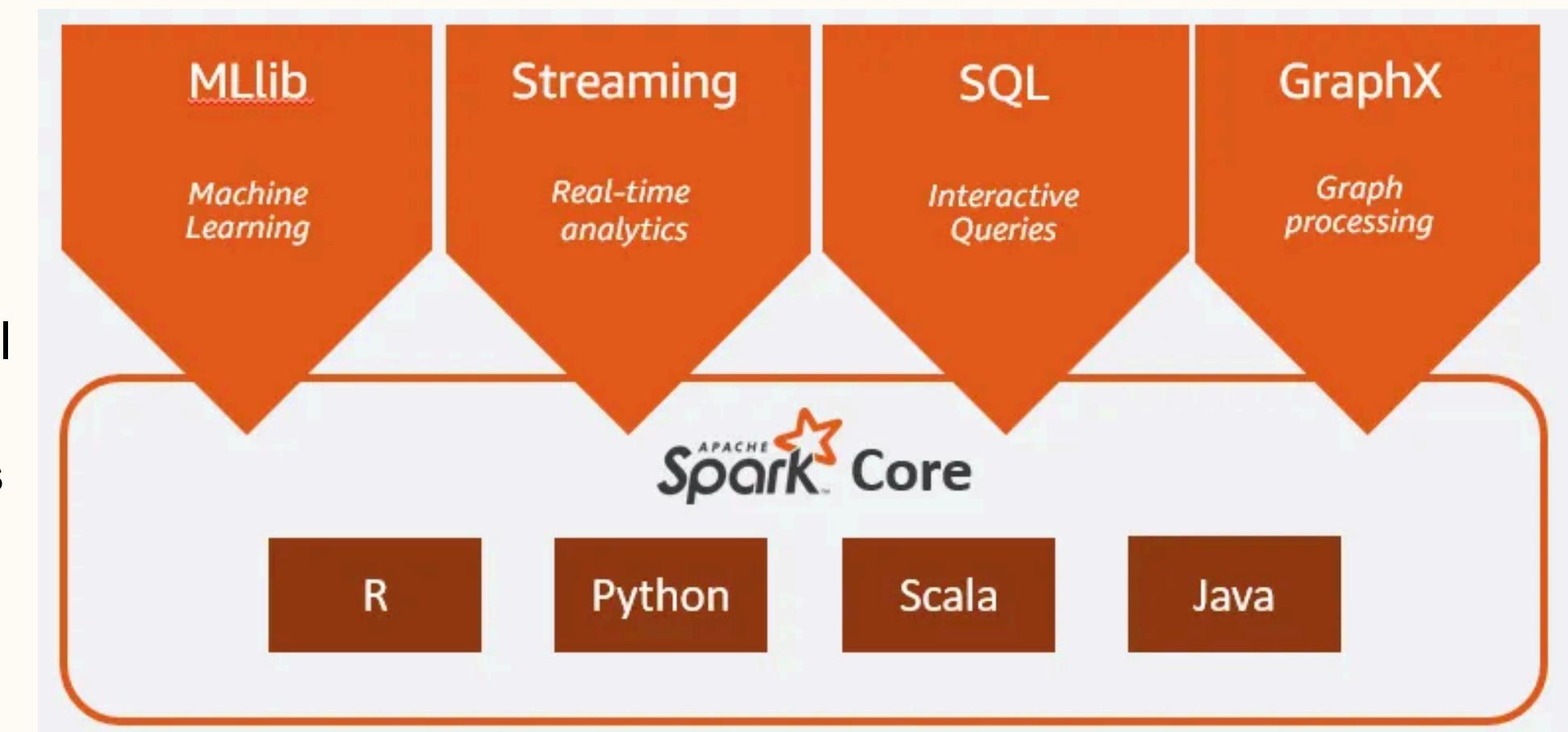


Pyspark: Computação Distribuída para Big Data

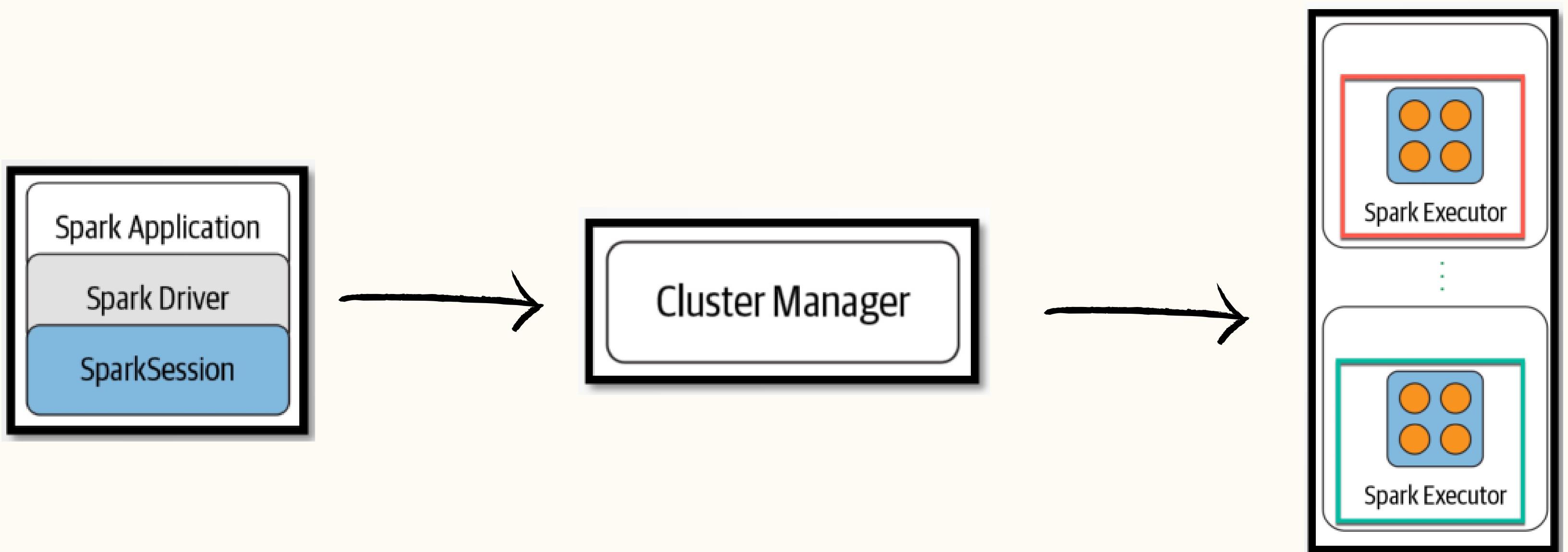


O Framework Spark

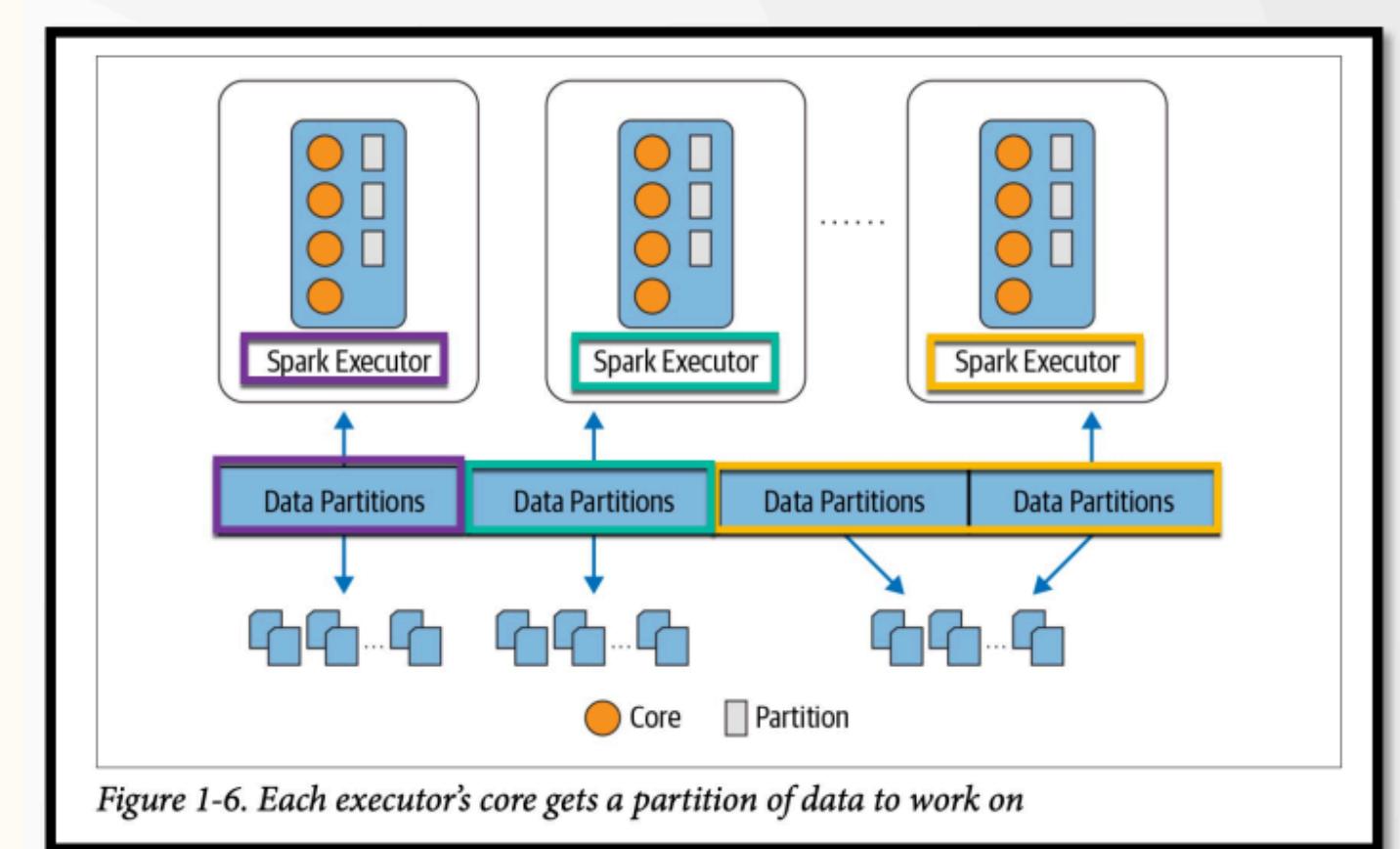
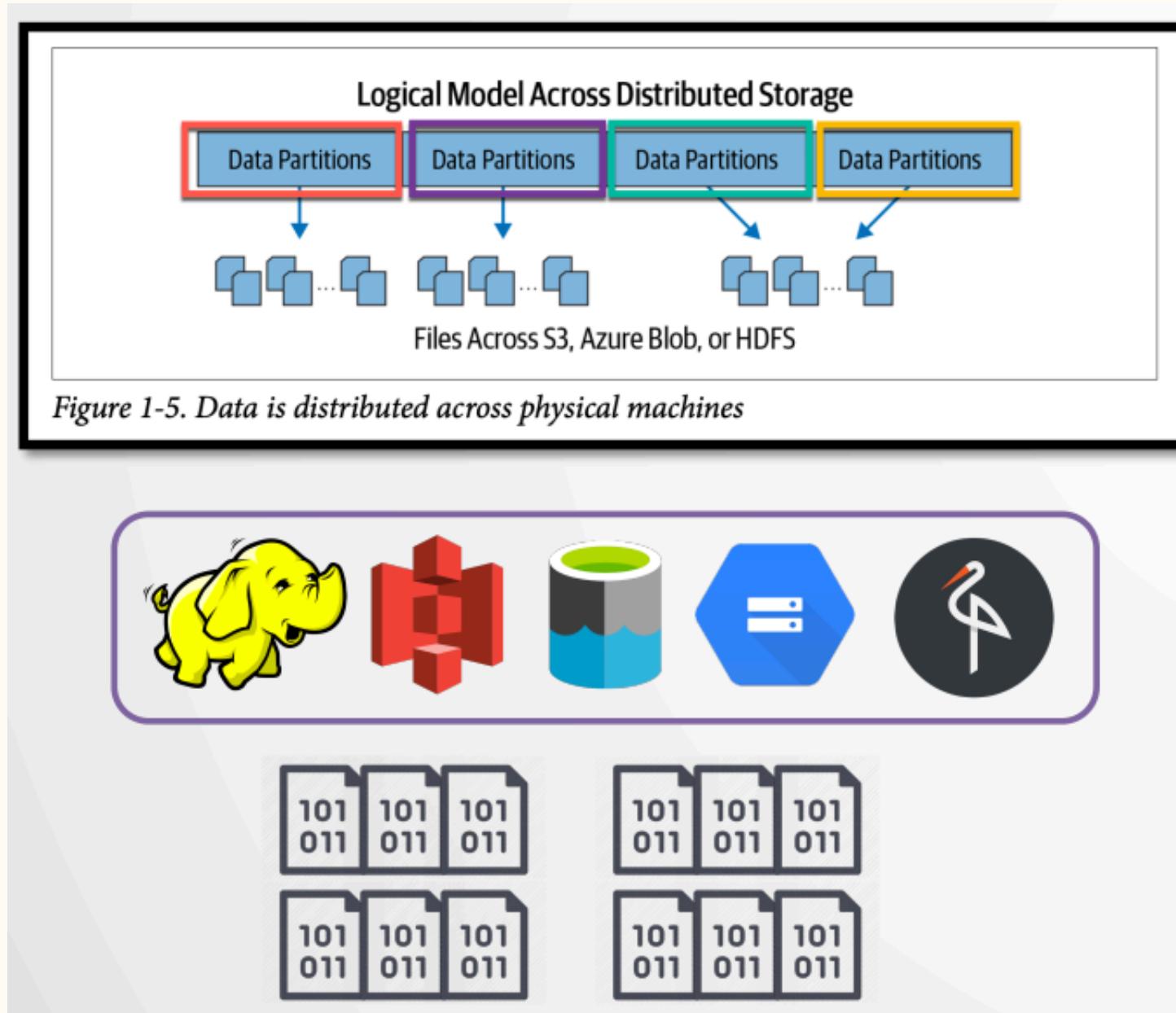
- **Spark Core** como a base da plataforma
- **Spark SQL** para consultas interativas
- **Spark Streaming** para análises em tempo real
- **Spark MLlib** para aprendizado de máquina
- **Spark GraphX** para processamento de grafos



Componentes Apache Spark



Processamento de Dados Distribuído



Hands On - Setup

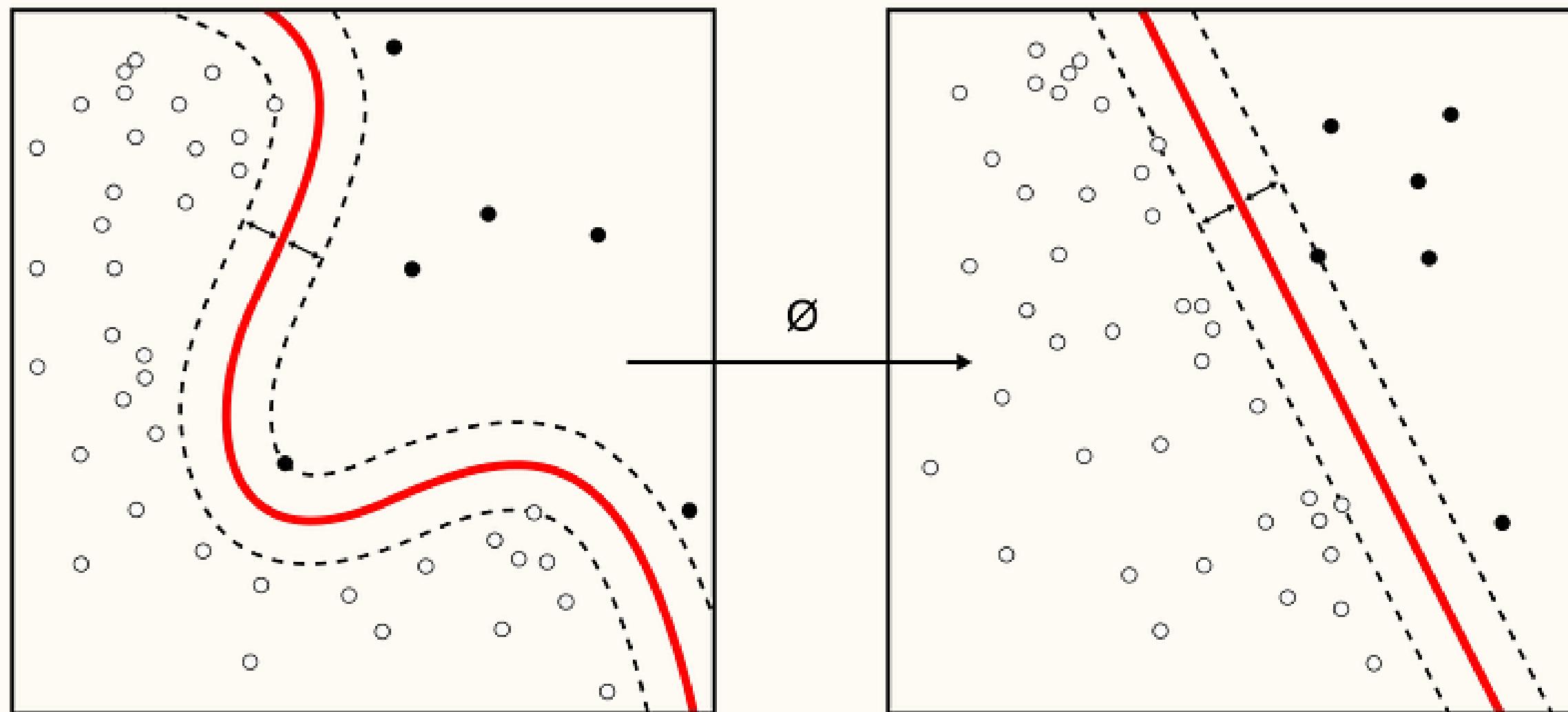
- 1.Crie uma conta Gmail, se ainda não tiver uma
- 2.Utilize sua conta do Google para criar um Colab Notebook
- 3.Crie uma conta no GitHub
- 4.Crie uma conta no Kaggle
- 5.Crie uma conta no Ngrok
- 6.Utilize a leitura de QR Code pra autenticação do Ngrok

Acesse o link: <http://bit.ly/3HnoUrZ>

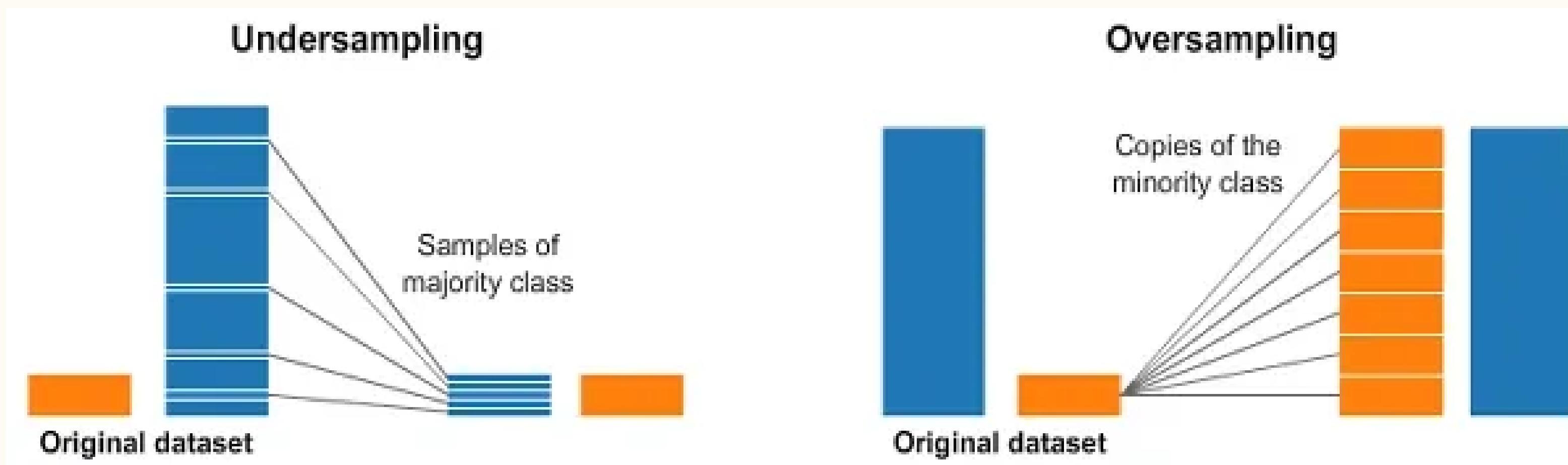
Case: Prevenção à Fraudes



Problemas de Classificação



Dados Desbalanceados





Especialidades em Dados

Explorando as Principais Áreas de Atuação

Conectando Tecnologia e Estratégia com Dados

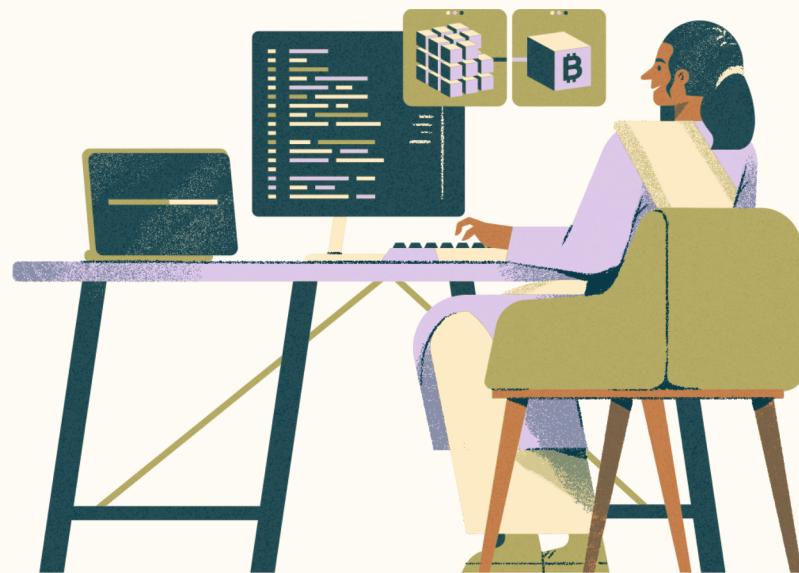
A área de dados é uma ponte essencial entre tecnologia e negócios, pois conecta a capacidade técnica de manipular e analisar dados com a necessidade estratégica de tomar decisões informadas.

Algumas carreiras em dados estão mais próximas da tecnologia, enquanto outras estão mais alinhadas aos negócios, criando um espectro de funções que trabalham juntas para gerar valor.

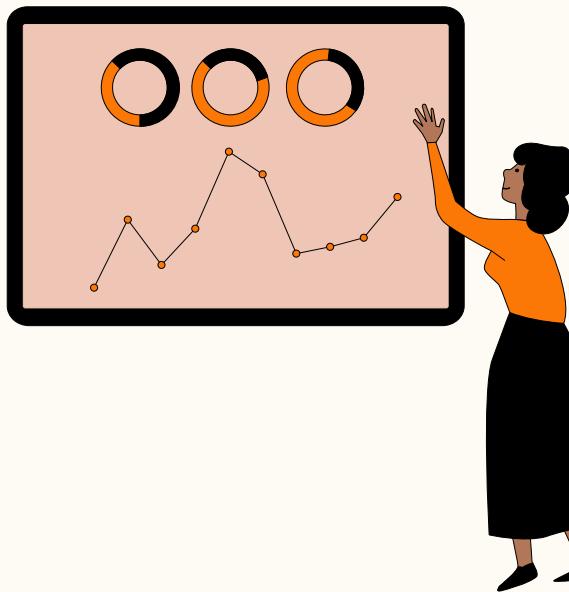


Principais Perfis Profissionais em Dados

Engenharia de Dados



Análise de Dados



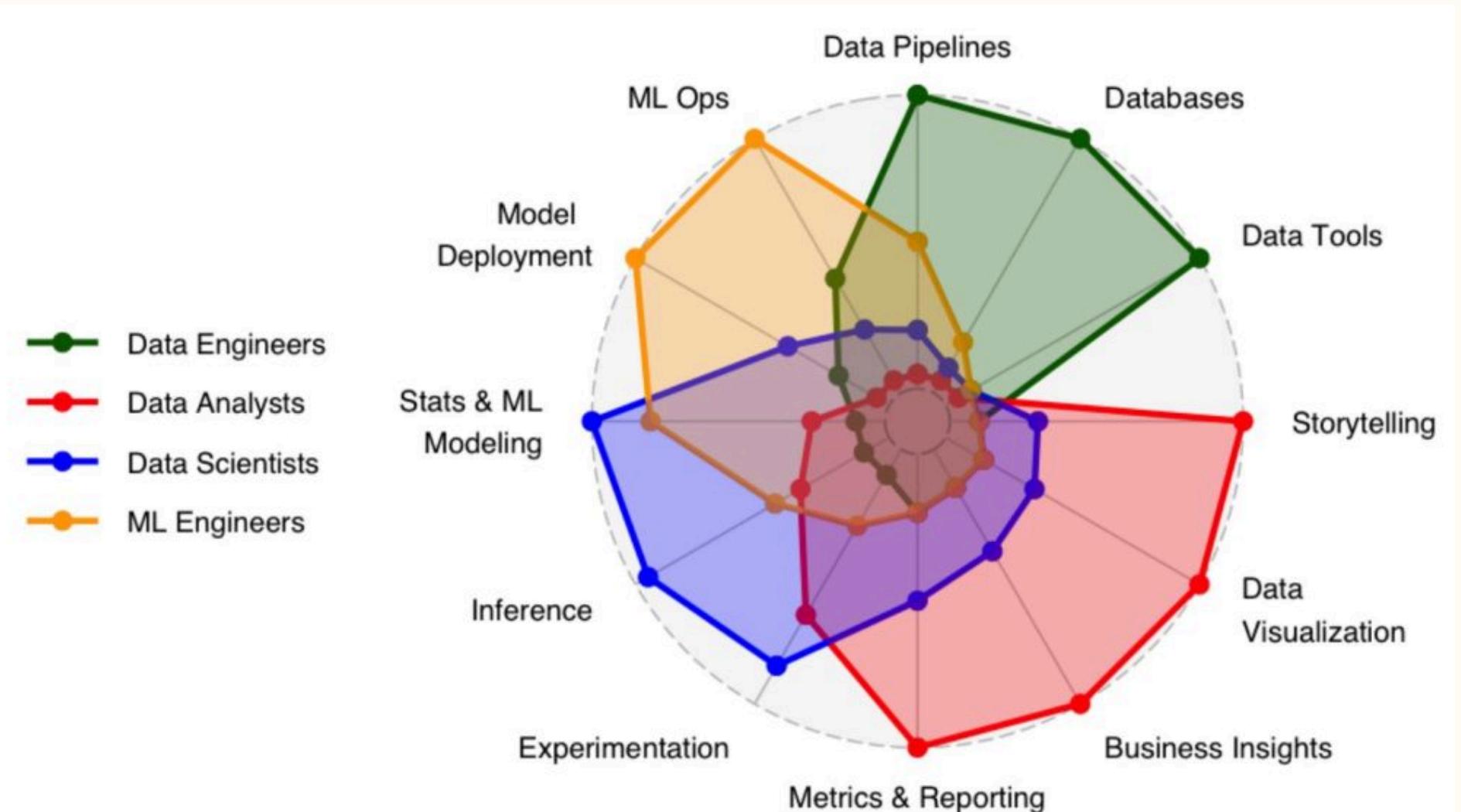
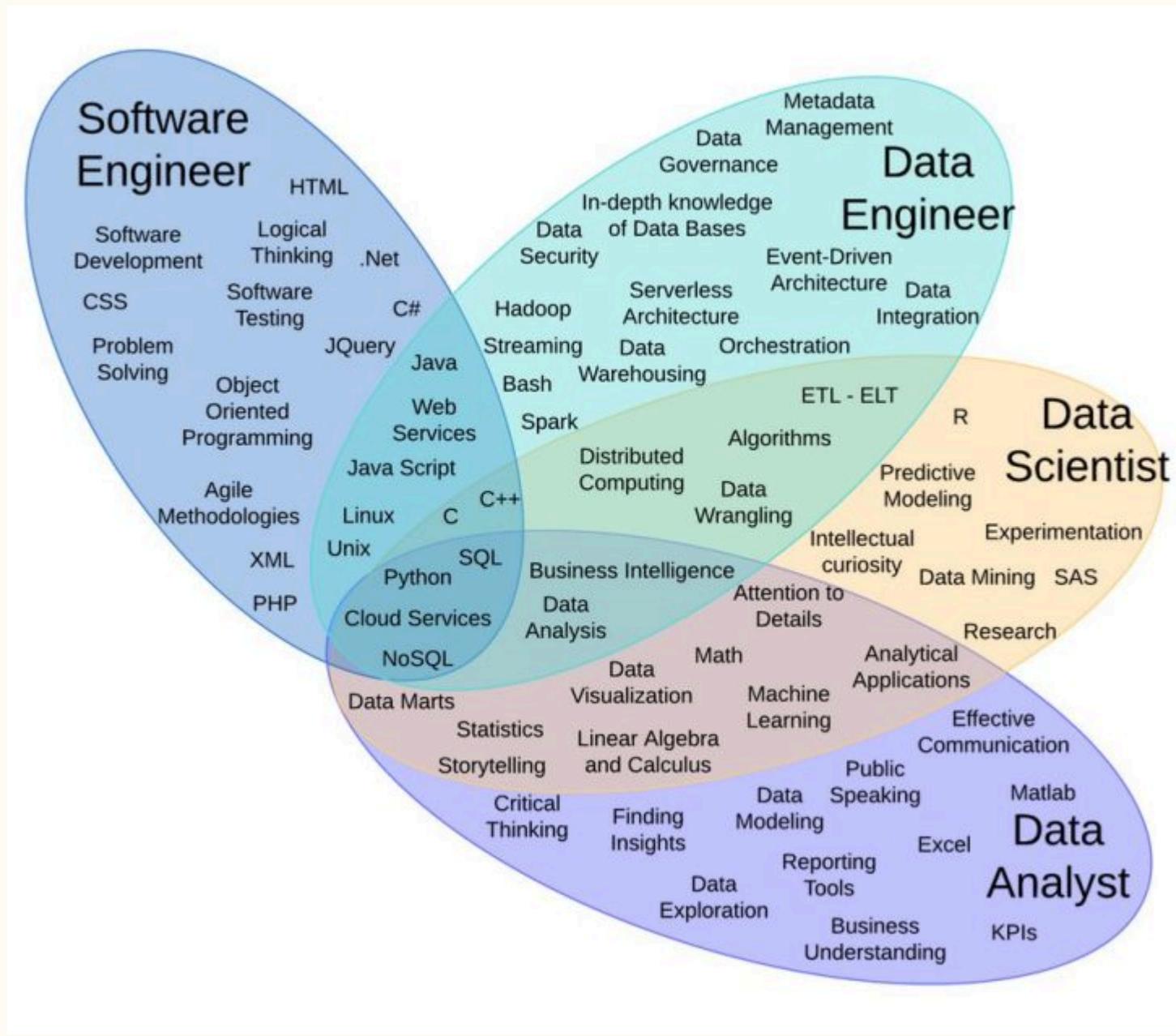
Engenharia de
Machine Learning



Ciência de Dados

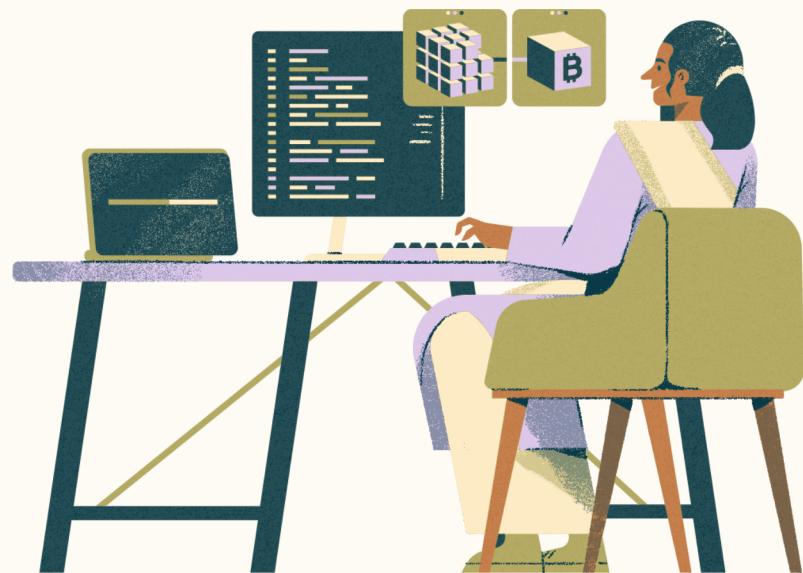


Mapa de Habilidades Técnicas



Fonte: Artigo "What's a data scientist? Explaining roles in big data", Hibernian Recruitment

Engenharia de Dados



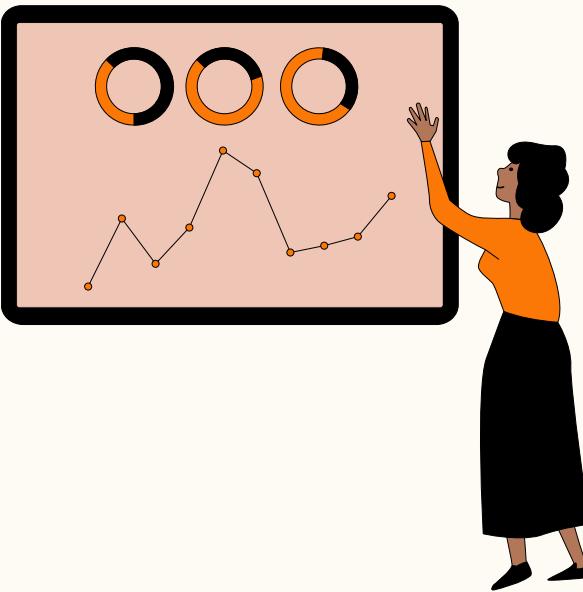
Responsabilidades:

- Infraestrutura: Desenvolve e mantém pipelines e arquiteturas de dados para garantir o fluxo, transformação e armazenamento eficiente.
- Preparação de Dados: Garante que os dados estejam limpos, estruturados e prontos para análise.
- Ferramentas: Trabalha com ETL, bancos de dados, plataformas em nuvem (AWS, Google Cloud, Azure) e frameworks de big data (Hadoop, Spark).
- Colaboração: Atua em parceria com cientistas de dados para fornecer conjuntos de dados utilizáveis para análise e modelagem.

Habilidades:

- Programação em Python, Java e SQL.
- Expertise em bancos de dados, data warehousing e computação em nuvem.
- Conhecimento em sistemas distribuídos e ferramentas como Apache Kafka e Apache Airflow.

Análise de Dados



Responsabilidades:

- Análise de Dados: Foca na interpretação de dados estruturados para identificar tendências e insights.
- Relatórios: Cria dashboards, relatórios e visualizações para comunicar os resultados.
- Alinhamento com Negócios: Trabalha em estreita colaboração com equipes de negócios para fornecer insights acionáveis para a tomada de decisões.

Habilidades:

- Proficiência em ferramentas de business intelligence (ex.: Tableau, Power BI).
- Forte base em estatística e visualização de dados.
- Conhecimento de SQL para consultas em bancos de dados estruturados.

Engenharia de Machine Learning



Responsabilidades:

- Implantação de Modelos: Implementa modelos de machine learning em sistemas de produção.
- Otimização: Garante a escalabilidade, desempenho e confiabilidade dos modelos implantados.
- Automação: Desenvolve sistemas que automatizam tarefas usando IA e aprendizado de máquina.

Habilidades:

- Fortes habilidades de programação em Python, Java ou R.
- Conhecimento de frameworks de machine learning (ex.: TensorFlow, PyTorch, Scikit-learn).
- Proficiência em ferramentas de MLOps (Machine Learning Operations) para gerenciamento do ciclo de vida dos modelos.

Ciência de Dados



Responsabilidades:

- Exploração de Dados: Analisa dados estruturados e não estruturados para derivar insights.
- Desenvolvimento de Modelos: Utiliza técnicas de aprendizado de máquina e estatística para criar modelos preditivos.
- Testes de Hipóteses: Formula e testa hipóteses para resolver problemas complexos de negócios.

Habilidades:

- Expertise em estatística, aprendizado de máquina e visualização de dados.
- Proficiência em linguagens de programação (Python, R) e ferramentas como Jupyter Notebooks.
- Forte habilidade para comunicar insights a partes interessadas técnicas e não técnicas.

Referências

- Domo (2024) – Infográfico “Data Never Sleeps”.
- Matt Turck (2024) - The 2024 MAD (Machine Learning, AI and Data) Landscape.
- [Documentação Oficial Spark](#)
- [Documentação Oficial Plotly](#)
- [Documentação Oficial Streamlit](#)
- [Documentação Oficial imblearn](#)
- [Artigo sobre Dados Desbalanceados em Casos de Detecção](#)
- [GitHub Developer Student Pack](#)
- [Bolsas Mestrado](#)
- [Bolsas Doutorado](#)

**“Todos precisam de alfabetização em dados,
porque os dados estão em todos os lugares. É a
nova moeda, é a linguagem dos negócios.
Precisamos ser capazes de falar essa
linguagem.”**

— Piyanka Jain

Contato



<https://qrco.de/bfmC8K>

Valeu!