Aula 02 - 09/09/2025

Conteúdo do Curso/Ferramentas que vamos aprender durante o curso

* -Consultas SQL
* -Revisão Python
* -Pyspark
* -Pandas
* -Matploilib

**SQL (Structured Query Language):** Linguagem de domínio específico utilizada para gerenciar dados em sistemas de gerenciamento de banco de dados relacionais (RDBMS), permitindo a execução de operações como definição de dados (DDL), manipulação de dados (DML) e controle de dados (DCL).

**Python:** Linguagem de programação de alto nível, interpretada e multiparadigma, amplamente utilizada em desenvolvimento web, análise de dados, inteligência artificial e automação, devido à sua sintaxe clara e vasta gama de bibliotecas.

**PySpark:** API Python para o Apache Spark, um framework de código aberto para processamento distribuído de grandes volumes de dados. O PySpark permite a manipulação de dados em larga escala utilizando o modelo de programação MapReduce, aproveitando a capacidade de processamento em memória do Spark.

**Pandas:** Biblioteca de código aberto para a linguagem Python, que oferece estruturas de dados flexíveis e de alto desempenho para manipulação e análise de dados. É amplamente utilizada para operações de limpeza, transformação, agregação e visualização de dados.

**Matplotlib:** Biblioteca de plotagem para a linguagem de programação Python e sua extensão numérica NumPy. Fornece uma API orientada a objetos para incorporar gráficos em aplicativos, permitindo a criação de uma ampla variedade de visualizações estáticas, animadas e interativas

**MapReduce:** é uma ferramenta para BigData que distribui o processamento de dados na memória RAM para carregar de forma mais leve. Essencialmente, ele mapeia os dados, os agrupa, e depois os reduz para um resultado final.

**vídeo visto na aula**

([Google Data Center 360° Tour](https://youtu.be/zDAYZU4A3w0?si=icbJtWAMYfDu8qJx))

**Site de dados de rio preto (violencia)**

(https://saude.riopreto.sp.gov.br/portal/boletim/iframe-viewer.html?boletim=violencias.html)

**Um bom projeto de dados deve conter**

**1. Coleta de Dados**

* **Explicação/Exemplo:** É o processo de adquirir dados de diversas fontes. Exemplo: Baixar um arquivo CSV de um site governamental, usar uma API para obter dados de redes sociais ou raspar dados de websites (web scraping).
* **Ferramentas:** Python (com bibliotecas como requests, BeautifulSoup, Scrapy), SQL (para extrair de bancos de dados), APIs específicas.

**2. Limpeza de Dados**

* **Explicação/Exemplo:** Tratar dados brutos para corrigir erros, remover duplicatas, lidar com valores ausentes ou inconsistências. Exemplo: Converter todos os nomes de cidades para maiúsculas para padronizar, remover linhas com idade zero, ou preencher valores nulos com a média.
* **Ferramentas:** Pandas (Python), SQL, Excel (para conjuntos de dados menores).

**3. Mineração de Dados**

* **Explicação/Exemplo:** Aplicar algoritmos para descobrir padrões, tendências e informações úteis em grandes conjuntos de dados. Exemplo: Identificar que clientes que compram o produto A também tendem a comprar o produto B (regras de associação), ou prever o churn de clientes.
* **Ferramentas:** Scikit-learn (Python), R, Spark (com PySpark), algoritmos de Machine Learning.

**4. Análise de Conteúdo**

* **Explicação/Exemplo:** Processar e interpretar informações de texto, imagens, áudio ou vídeo para extrair insights. Geralmente se refere à análise de dados não estruturados. Exemplo: Analisar sentimentos em comentários de clientes para entender a percepção de um produto, ou categorizar notícias por tema.
* **Ferramentas:** NLTK (Natural Language Toolkit) (Python), spaCy (Python), bibliotecas de Visão Computacional (OpenCV), ferramentas de Processamento de Linguagem Natural (PLN).