PROJETO APLICADO DE PIPELINE DE DADOS - FASE 01

Disciplina: Pipeline de Dados - IA-CD

VISÃO GERAL DO PROJETO

Vocês irão desenvolver um **pipeline de dados completo e funcional** que resolva um problema real de negócio. Este projeto será desenvolvido em **4 fases** ao longo do semestre, permitindo que vocês apliquem gradualmente os conceitos aprendidos em aula.

Por que este projeto é importante?

- **Experiência real:** Vocês trabalharão com dados reais e problemas autênticos
- **Portfolio:** Projeto completo para mostrar para empresas
- **Aprendizado prático:** Aplicação direta dos conceitos teóricos
- **Trabalho em equipe:** Simulação de ambiente profissional

O QUE VOCÊS JÁ SABEM (AULAS 01-04)

Com base nas aulas anteriores, vocês já dominam:

Fundamentos de Pipeline

- Etapas: Coleta → Transformação → Armazenamento → Análise
- Conceitos de ETL (Extract, Transform, Load)
- Importância da qualidade dos dados

Python & Pandas

- Manipulação de DataFrames
- Limpeza de dados (`fillna()`, `dropna()`, `drop()`)
- Transformações básicas (criar colunas, filtros, agrupamentos)
- Leitura de arquivos CSV

Bancos de Dados

- SQLite para desenvolvimento
- Conceitos de MySQL e PostgreSQL
- Operações básicas de SQL
- Integração Python-Banco

Apache Spark (Introdução)

- Diferenças entre Spark e Pandas
- DataFrames distribuídos
- Processamento de Big Data
- Databricks Serverless

O QUE AINDA VAMOS APRENDER

Nas próximas aulas, vocês aprenderão ferramentas que se integrarão ao projeto:

Orquestração de Pipelines

- Apache Airflow
- Automação de workflows
- Agendamento de tarefas

Cloud Computing

- AWS/Azure/GCP para Big Data
- Data Lakes e Data Warehouses
- Infraestrutura como Código (IaC)

Visualização e BI

- Dashboards interativos
- Power BI / Tableau
- Storytelling com dados

ML Ops

- Pipelines para Machine Learning
- Feature Engineering automatizado
- Modelos em produção

FASE 01: ESCOLHA DO TEMA E PLANEJAMENTO

Objetivo da Fase 01:

Definir o problema de negócio, escolher o dataset e criar o plano inicial do pipeline.

TEMAS DISPONÍVEIS PARA ESCOLHA

Escolham **UM** dos temas abaixo. Cada tema foi pensado para permitir evolução gradual:

1. E-COMMERCE ANALYTICS

Problema: Analisar comportamento de compras para otimizar vendas
Datasets sugeridos:

- E-commerce Data
- Online Retail

Pipeline inicial: Vendas → Limpeza → Análise de padrões → Dashboard
Evolução futura: Recomendações em tempo real, previsão de demanda

---

2. ANÁLISE DE REDES SOCIAIS

Problema: Monitorar sentimentos e tendências em redes sociais
Datasets sugeridos:

- Twitter Sentiment
- Social Media Posts

Pipeline inicial: Posts → Limpeza de texto → Análise de sentimento → Visualização
Evolução futura: Processamento em tempo real, alertas automáticos

---

3. DADOS DE SAÚDE

Problema: Analisar dados médicos para insights de saúde pública
Datasets sugeridos:

- Heart Disease
- COVID-19 Data

Pipeline inicial: Dados médicos → Limpeza → Análise epidemiológica → Relatórios
Evolução futura: Predição de surtos, dashboard em tempo real

---

4. MOBILIDADE URBANA

Problema: Otimizar transporte urbano e reduzir trânsito
Datasets sugeridos:

- NYC Taxi Data
- Uber/Lyft Data

Pipeline inicial: Viagens → Limpeza → Análise de padrões → Mapas
Evolução futura: Previsão de demanda, otimização de rotas

---

5. ANÁLISE FINANCEIRA

Problema: Detectar padrões e anomalias em transações financeiras
Datasets sugeridos:

- Credit Card Fraud
- Stock Market Data

Pipeline inicial: Transações → Limpeza → Detecção de anomalias → Alertas
Evolução futura: ML para fraude, processamento em tempo real

---

6. AGRONEGÓCIO E SUSTENTABILIDADE

Problema: Otimizar produção agrícola usando dados ambientais
Datasets sugeridos:

- Crop Production
- Weather Data

Pipeline inicial: Dados agrícolas → Limpeza → Análise de produtividade → Insights
Evolução futura: IoT sensors, previsão de safras

ENTREGÁVEIS DA FASE 01

1. DOCUMENTO DE PROPOSTA (2-3 páginas)

Criem um documento com:

1.1 Identificação do Grupo

- Nomes dos integrantes
- Tema escolhido
- Justificativa da escolha

1.2 Definição do Problema

- Qual problema de negócio vocês querem resolver?
- Por que este problema é importante?
- Qual o impacto esperado da solução?

Exemplo para E-commerce:

> *"Nosso e-commerce fictício está perdendo clientes e não sabe por quê. Queremos criar um pipeline que analise dados de vendas, identifique padrões de abandono de carrinho e sugira estratégias para aumentar as conversões."*

1.3 Descrição dos Dados

- Qual dataset escolheram?

- Quantas linhas e colunas tem?
- Quais são as principais variáveis?
- Que problemas de qualidade identificaram?

1.4 Arquitetura Inicial do Pipeline

Desenhem um diagrama simples mostrando:

```
[Fonte de Dados] → [Extração] → [Transformação] → [Armazenamento] → [Análise]
```

2. ANÁLISE EXPLORATÓRIA INICIAL (Jupyter Notebook)

Criem um notebook com:

2.1 Carregamento dos Dados

```
import pandas as pd
import matplotlib.pyplot as plt

# Carregar dataset
df = pd.read_csv('seu_dataset.csv')
print(f"Dataset carregado: {df.shape}")
```

2.2 Exploração Básica

- `df.info()`
- `df.describe()`
- `df.isnull().sum()`
- Primeiras visualizações

2.3 Identificação de Problemas

- Valores faltantes
- Duplicatas
- Outliers

- Inconsistências

2.4 Primeiras Transformações

Apliquem pelo menos 3 transformações:

- Limpeza de valores nulos
- Criação de uma nova coluna
- Filtro de dados relevantes

3. PLANEJAMENTO DAS PRÓXIMAS FASES

Fase 02 - Transformação Avançada (Próxima entrega)

- Quais transformações complexas farão?
- Como vão integrar múltiplas fontes?
- Que métricas de qualidade implementarão?

Fase 03 - Big Data e Cloud (Futuro)

- Como migrarão para Spark?
- Que serviços de cloud usarão?
- Como automatizarão o pipeline?

Fase 04 - Produção e ML (Final)

- Que dashboards criarão?
- Implementarão machine learning?
- Como colocarão em produção?

CRONOGRAMA E ENTREGAS

Fase 01 - Planejamento (ATUAL)

- **Entrega:** 13/10/2025
- **Formato:** Documento PDF + Notebook

Próximas Fases:

- **Fase 02:** Transformação Avançada + Bancos
- **Fase 03:** Big Data + Cloud + Automação
- **Fase 04:** Produção + Dashboard Final

DICAS DO PROFESSOR

Foco na Simplicidade (Por Enquanto)

- Comecem simples e evoluam gradualmente
- Usem as ferramentas que já conhecem (Python + Pandas)
- Não se preocupem com Big Data ainda

Escolham Dados Interessantes

- Dados que contem uma história
- Datasets com problemas reais para resolver
- Volume suficiente para análises significativas

Trabalho em Equipe

- Dividam as tarefas de forma equilibrada
- Usem Git/GitHub para colaboração
- Documentem tudo que fazem

Pensem no Futuro

- Escolham temas que permitam evolução
- Considerem onde querem chegar na Fase 04
- Mantenham flexibilidade para mudanças

"Todo grande pipeline começou com um simples `pd.read_csv()` *"*

Boa sorte e mãos à obra!