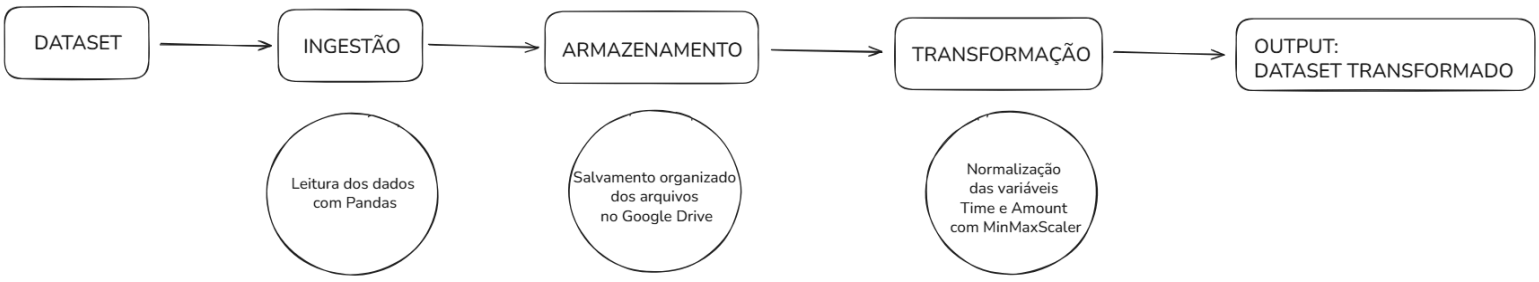
Projeto em Fundamentos de Big Data

Nome dos(as) Alunos(as): Arthur Lins, Lucas Fernandes e Victor Aroucha
Título do Projeto/Demanda: Detecção de transações fraudulentas com cartão de crédito em tempo real

AV1:

Diagrama do Pipeline de Dados:



Tecnologias Utilizadas e Possíveis Refinamentos:

Ingestão: Python + Pandas
Armazenamento: Google Drive
Transformação: Scikit-learn (MinMaxScaler)
Visualização: Matplotlib / Seaborn

Tecnologias que poderiam ser usadas para refinamento (pagas ou avançadas):

Ingestão em tempo real: Apache Kafka (Permitem ingestão de dados em tempo real em grande escala)
Armazenamento distibuído: AWS ou Databricks (Armazenamento em nuvem de alta disponibilidade e integração com pipelines automatizados)
Processamento de grandes volumes: Apache Spark (Databricks) (Escalabilidade e processamento paralelo de grandes datasets)
Visualização: Power BI (Dashboards interativos e monitoramento em tempo real)

Arquitetura parcial implementada (ambiente simulado, processamento em batch):

Visão geral: Implementamos um ETL simples em batch no Google Colab, usando o Google Drive como armazenamento. O fluxo foi: ingestão (leitura completa do CSV), transformação (normalização), armazenamento do resultado.

Ingestão (batch)

- Leitura direta do arquivo creditcard.csv a partir do Drive: /content/drive/MyDrive/BigData/creditcard.csv
- Checagens rápidas: df.shape, df.info(), verificação de nulos (df.isnull().sum() = 0), distribuição da coluna alvo (df['Class'].value_counts()) e inspeção das 5 primeiras fraudes (df[df['Class']==1].head())

Armazenamento

- Organização dos artefatos no Drive sob /content/drive/MyDrive/BigData

Transformação

- As colunas **Time** e **Amount** foram **normalizadas** com **MinMaxScaler** (intervalo [0, 1]).
- As demais variáveis (V1…V28) já estão padronizadas (derivadas de PCA) e **não precisaram de ajuste**
- Visualizações: histogramas de Amount e Time **antes e depois** da normalização para justificar a escolha (diferença de escala e cauda longa em Amount)

Decisões de design

- **Batch** em vez de streaming, por ser a entrega da **AV1** (foco em ingestão, transformação e armazenamento)
- **MinMaxScaler** escolhido para alinhar a escala de Time e Amount às demais features (PCA), sem alterar a forma da distribuição

Equipe responsável e divisão de tarefas:

INTEGRANTE	RESPONSABILIDADES PRINCIPAIS
Arthur Lins	Ingestão (leitura completa do CSV), transformação (normalização), visualizações e documentação no Colab.
Lucas Fernandes	Organização do Github, estrutura do repositório e revisão do Documento de Arquitetura.
Victor Aroucha	Consolidação das justificativas técnicas (PCA, normalização), diagrama do pipeline e revisão final do material da AV1.