O projeto implanta um pipeline reprodutível para um recomendador de filmes a partir de movies.csv e ratings.csv, organizado em três camadas. Na bronze, apenas ingerimos e padronizamos os dados brutos em Parquet para leitura eficiente e reprocesso. Na silver, executamos a transformação crítica: tipagem (datas e numéricos), normalização de campos textuais/JSON (ex.: genres), tratamento de nulos e deduplicação; também “explodimos” a relação filme–gênero para viabilizar agregações. Na gold, publicamos artefatos prontos ao consumo: um quadro por filme com weighted rating (média ponderada por volume de votos, combinada à popularidade) e perfis simples de preferência por gênero por usuário, base para o baseline de recomendação.

A ingestão será batch (com opção didática de micro-lotes para simular streaming), a transformação usa Python + Pandas/NumPy e o armazenamento colunar PyArrow/Parquet. A demonstração ocorre em Jupyter/Colab, com amostras e gráficos rápidos (Matplotlib/Plotly). Para escala futura, tecnologias pagas (BigQuery, Snowflake ou lakehouse gerenciado) permanecem como evolução natural, mantendo nesta AV1 o cenário local e transparente.

A equipe se divide em três frentes: Dados/ETL (modelagem de bronze/silver/gold e critérios de qualidade), Modelagem/Recomendação (baseline que combina weighted rating e afinidade por gênero, com métricas iniciais), e Documentação/Repo (README, checklist, registros de execução e rastreabilidade). Essa estrutura cumpre a primeira entrega: pipeline iniciado, camadas operantes em ambiente simulado e justificativa tecnológica clara.