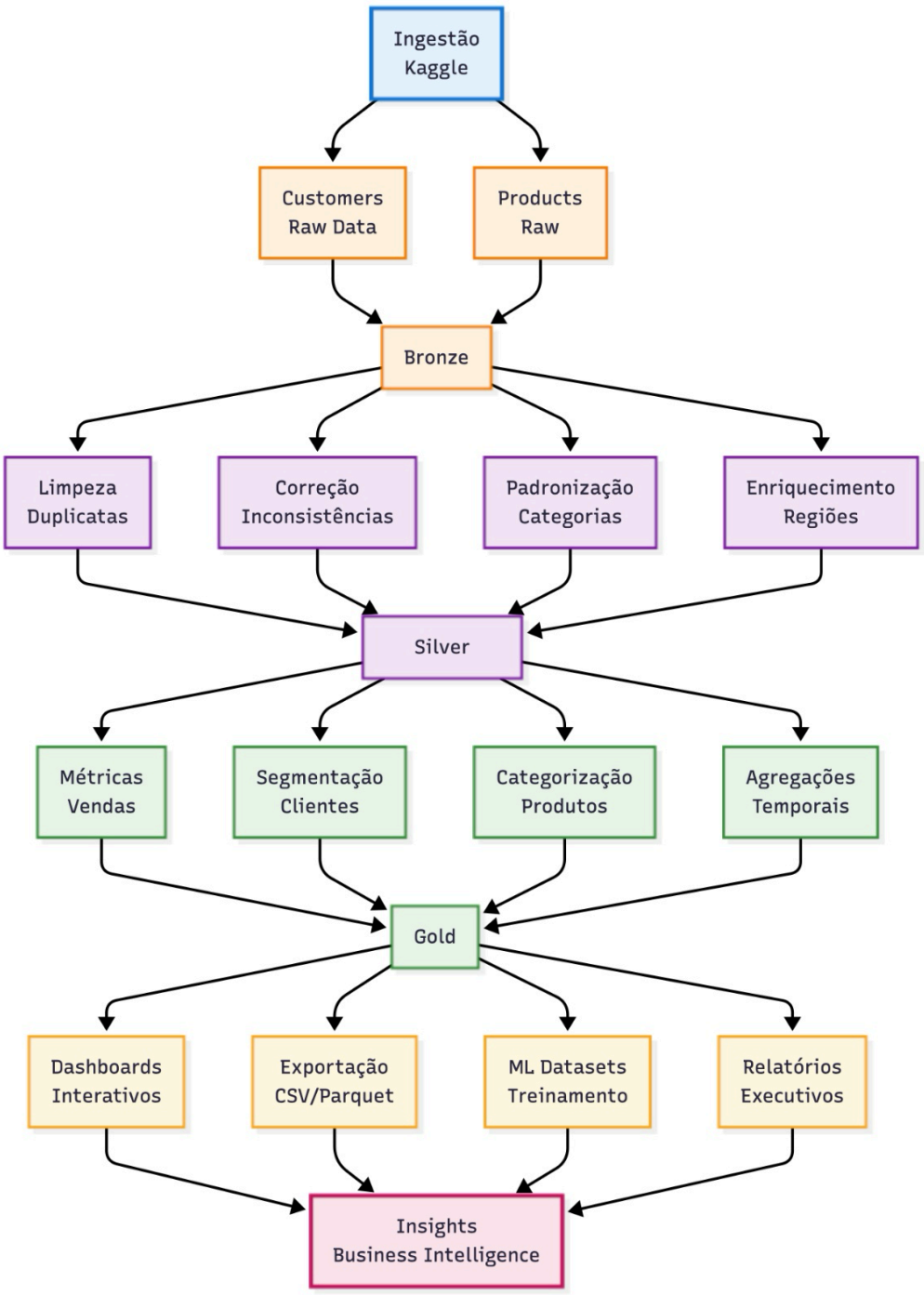
Arquitetura Medallion para Dados da Olist

1. Diagrama do Pipeline de Dados



2. Tecnologias Utilizadas e Possíveis Refinamentos

Ferramentas Utilizadas (gratuitas / open source)

Categoria	Ferramenta	Finalidade
Linguagem	Python 3.x	Base do pipeline e análise
Manipulação de dados	Pandas / NumPy	Limpeza, transformação e cálculos
Visualização	Matplotlib / Seaborn / Plotly	Geração de gráficos estáticos e interativos
Ambiente	Google Colab	Execução dos códigos e visualização de resultados
Armazenamento	GitHub	Registro do notebook com outputs e documentação
Documentação	Jupyter Notebook (.ipynb)	Organização de etapas e prints de execução

Todos os resultados e prints do pipeline foram gerados diretamente no Colab e salvos no notebook exportado ao GitHub.

Tecnologias Pagas (para refinamento futuro)

Categoria	Ferramenta	Justificativa
Orquestração de Pipeline	Apache Airflow (Cloud Composer - GCP)	Permite automatizar as etapas Bronze → Silver → Gold
Armazenamento Escalável	Google BigQuery	Otimiza consultas SQL sobre grandes volumes
Dashboards Profissionais	Tableau / Power BI Pro	Criação de painéis interativos e executivos
Monitoramento e Logs	Grafana Cloud / Datadog	Acompanha desempenho e falhas em pipelines reais
Machine Learning em Nuvem	Vertex AI (GCP)	Facilita treinamento e versionamento de modelos preditivos

Essas ferramentas poderiam tornar o pipeline automatizado e escalável, adequando-o a um contexto real de Big Data corporativo.

3. Arquitetura Parcial Implementada

O projeto foi desenvolvido em um ambiente acadêmico simulado, com execução manual no Google Colab.
As camadas foram aplicadas conforme a Arquitetura Medallion, com dados públicos da Olist:

Camada	**Implementação**	**Formato**
Bronze	Upload manual de arquivos CSV originais do Kaggle para o Colab	.csv
Silver	Limpeza e enriquecimento (remoção de duplicatas, padronização, criação de novas features)	.csv / .parquet
Gold	Análises, estatísticas, dashboards e prints de resultados exibidos no notebook	.ipynb

Os arquivos finais (prints, tabelas e gráficos) foram mantidos dentro do notebook e publicados no GitHub apenas como registro de execução.

4. Equipe Responsável e Divisão de Tarefas

Integrante	**Responsabilidade**
Arthur Hendrich Alencar de Menezes	Desenvolvimento da pipeline e criação de visualizações
Henrique Cordeiro Pereira	Desenvolvimento da pipeline, organização do repositório no GitHub e documentação no Collab.
Luiza Omena Suassuna	Desenvolvimento da pipeline, documentação no Colab e organização das demais documentações.

Projeto realizado como parte da disciplina de Big Data na CESAR School.