

# Arquitetura de Data Lake- CERC

Diagrama de arquitetura de um DataLake

## Origem dos dados



1

## Fonte de dados

As fontes de dados utilizadas no processo podem variar bastante. Exemplo: APIs, arquivos (CSV/ Parquet), bancos de dados e dados em nuvem



2

## Orquestração

5

## Processamento e Modelagem de Dados

Armazenamento escalável

### Bronze

Camada com os dados armazenados em seu formato mais bruto, exatamente como foram recebidos, sem nenhum tipo de transformação aplicada. Exemplo de diretórios e arquivos para o armazenamento Bronze

```
registros/cliente_x/ano=2025/.../dia=1/001.part
registros/cliente_x/ano=2025/.../dia=1/002.part
registros/cliente_x/ano=2025/.../dia=2/001.part
registros/cliente_x/ano=2025/.../dia=2/002.part
...
registros/cliente_z/ano=2025/.../dia=1/001.part
```

### Prata

Camada com dados já limpos, estruturados e filtrados. Colunas desnecessárias podem ser removidas, registros duplicados tratados e os dados estão prontos para análises exploratórias ou como base para agregações. Exemplos de transformações na camada Prata:

- Remoção de colunas irrelevantes dos registros
- Deduplicação de dados dimensionais
- Conversão e padronização de formatos de data ou identificadores

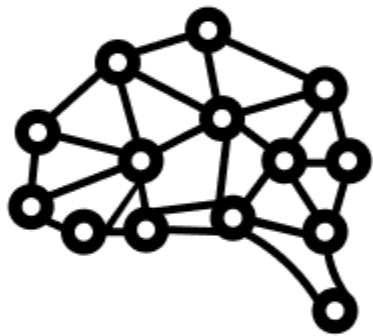
### Ouro

Camada final com dados otimizados para consumo. Os dados estão prontos para alimentar modelos de machine learning, visualizações analíticas e relatórios de negócio. Exemplo de dados ouro:

- Delta Tables fato com features para treinamento de modelos de Machine Learning

## Destino

Modelo de Machine Learning



4

## Simulação das transformação de dados para modelos de ML

Fluxo simples com duas fontes de dados e uma estrutura de tabelas fato e dimensão, passando pelas camadas Bronze, Prata e Ouro

