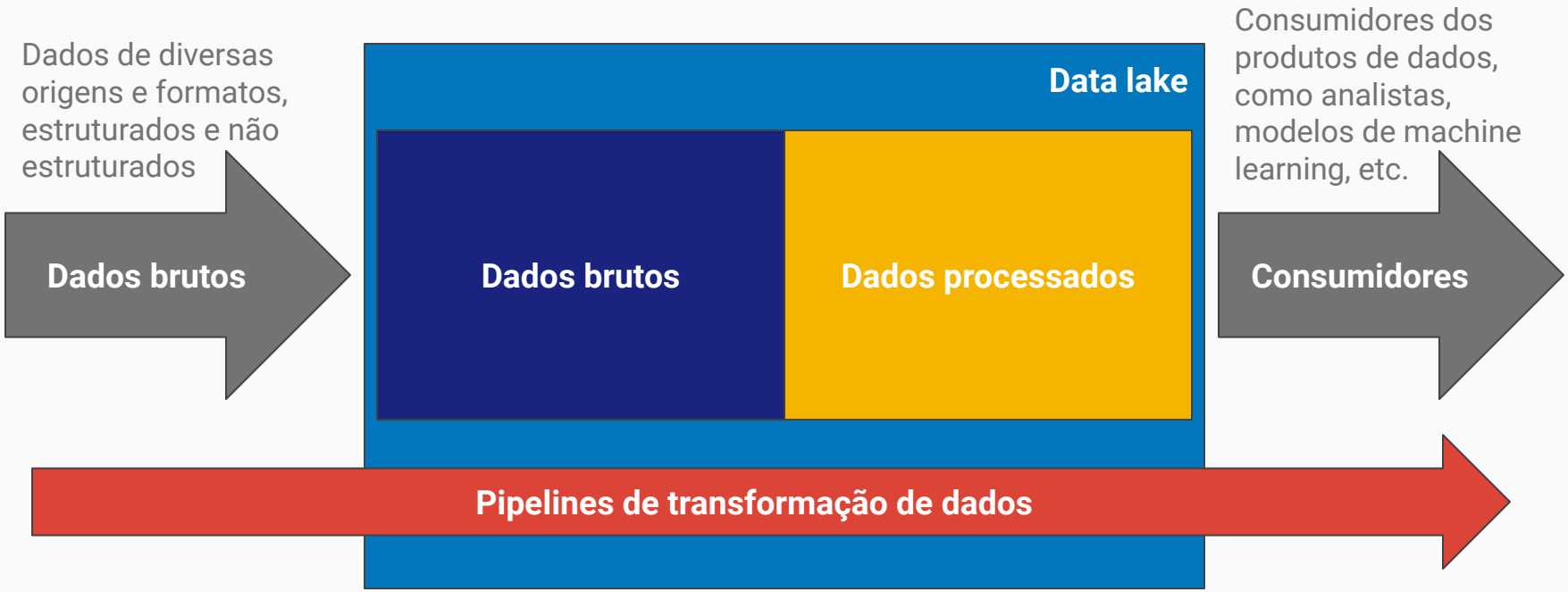
# Big Data

Tech Challenge - Fase 3 - Data Analytics


Dados de diversas origens e formatos, estruturados e não estruturados
Dados brutos
Data lake
Dados brutos
Dados processados
Consumidores dos produtos de dados, como analistas, modelos de machine learning, etc.
Consumidores
Pipelines de transformação de dados

# Dados brutos - origem

**Obtenção dos dados brutos em formato CSV:**

- Os dados foram carregados à partir de arquivos CSV obtidos no site do IBGE.
- Como o tema módulo é Big Data os dados foram manipulados utilizando Spra, através da SparkSession, o que permite a manipulação eficiente de grandes volumes de dados.

**Enriquecimento dos dados com informações de domínios dos campos:**

- A base de dados foi enriquecida utilizando dicionários auxiliares, como os obtidos do site do IBGE, para associar dados como estados (UF) e capitais, agregando valor aos dados brutos com informações de domínio relevantes.

# Data lake - origem

**Adoção do formato "wide table":**

- A escolha pelo formato "wide table" reflete uma prática comum em cenários de big data, onde todos os atributos relevantes são representados em uma única tabela. Isso facilita consultas e operações analíticas, melhorando a performance ao evitar a necessidade de joins complexos.

**Gravação da tabela com os dados brutos enriquecidos no data lake:**

- Após o enriquecimento dos dados, a tabela foi salva em um data lake (no caso, utilizando o BigQuery). Data lakes permitem o armazenamento de dados em sua forma bruta, sendo posteriormente acessíveis para tratamento ou análise.

# Data lake - tratamento dos dados

**Seleção de colunas mais significativas para análises:**

- Para tornar as análises mais eficientes, foi feita a seleção das colunas mais relevantes. As colunas foram renomeadas para nomes mais significativos e fáceis de entender, facilitando o trabalho dos analistas de dados que consumirão as tabelas.

**Gravação de tabelas no data lake e explicação das camadas:**

- O processo segue a estrutura de um data lake, onde os dados são armazenados em diferentes camadas: dados brutos (raw data) e dados tratados (processed data). Essa separação permite o versionamento e facilita o acesso a dados em diferentes estágios de tratamento.