Pipeline de Dados

O QUE É PIPELINE DE DADOS?

"Um pipeline de dados é uma forma de movimentar os dados de um local (origem) para um destino."

"É uma sequência de etapas de coleta, armazenamento, transformação para preparar dados corporativos para análise"

IMPORTANTE

A construção e a manutenção de um pipeline de dados é responsabilidade do Engenheiro de Dados.

Pipeline de Dados

COMPONENTES DE UM PIPELINE DE DADOS

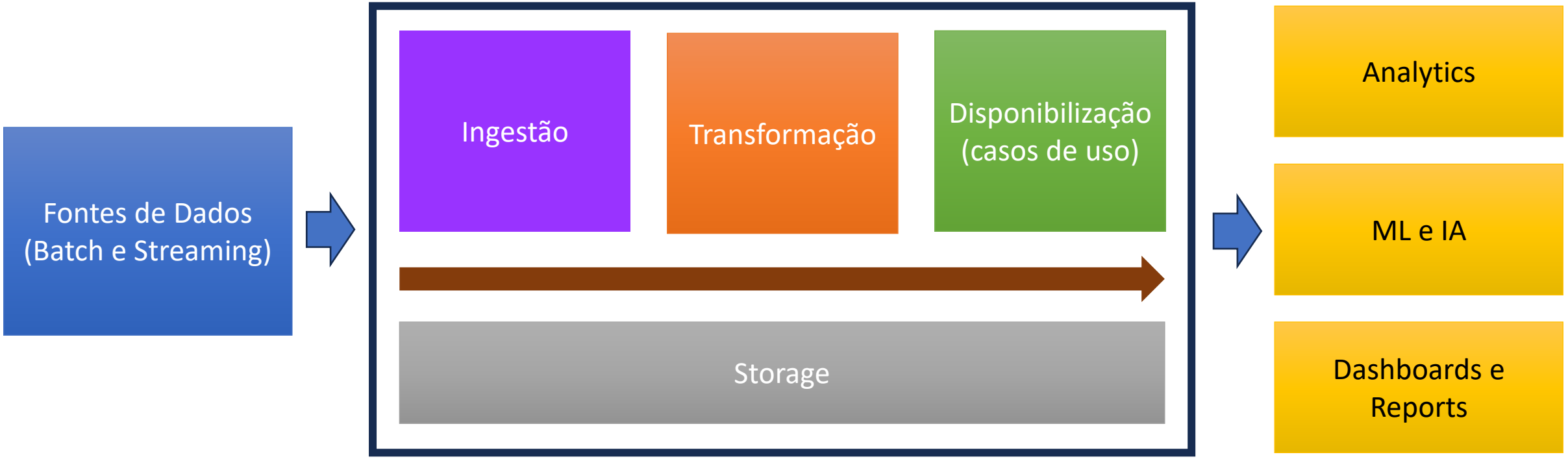
ORIGEM **PROCESSAMENTO** **DESTINO**

Só isso? ->

Fonte da imagem: Fundamentos de Engenharia de Dados: Projete e Construa Sistemas de Dados Robustos

Pipeline de Dados

COMPONENTES DE UM PIPELINE DE DADOS



Arquitetura de Dados

Gerenciamento de Dados e Metadados

Segurança

Engenharia de Software

Segurança

DataOps

Fonte da imagem: Fundamentos de Engenharia de Dados: Projete e Construa Sistemas de Dados Robustos

Pipeline de Dados
ETL X ELT

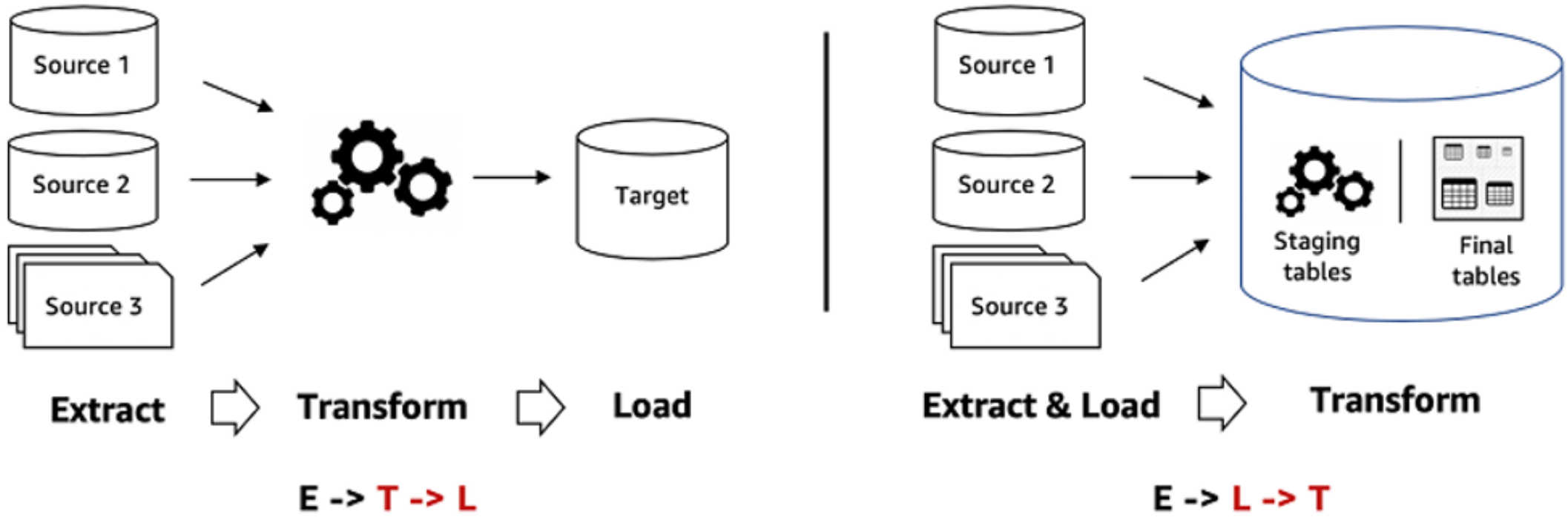
O processo de **extração de dados** é uma forma comum de combinação dos dados de vários sistemas em um único banco de dados, repositório de dados, armazenamento de dados ou data lakes.

Existem 2 abordagens para essa etapa de extração:

- **E T L -** Extract, Transform, Load (Extrair, Transformar, Carregar)
- **E L T -** Extract, Load, Transform (Extrair, Carregar, Transformar)

Fonte: https://cloud.google.com/learn/what-is-etl?hl=pt-br#section-1

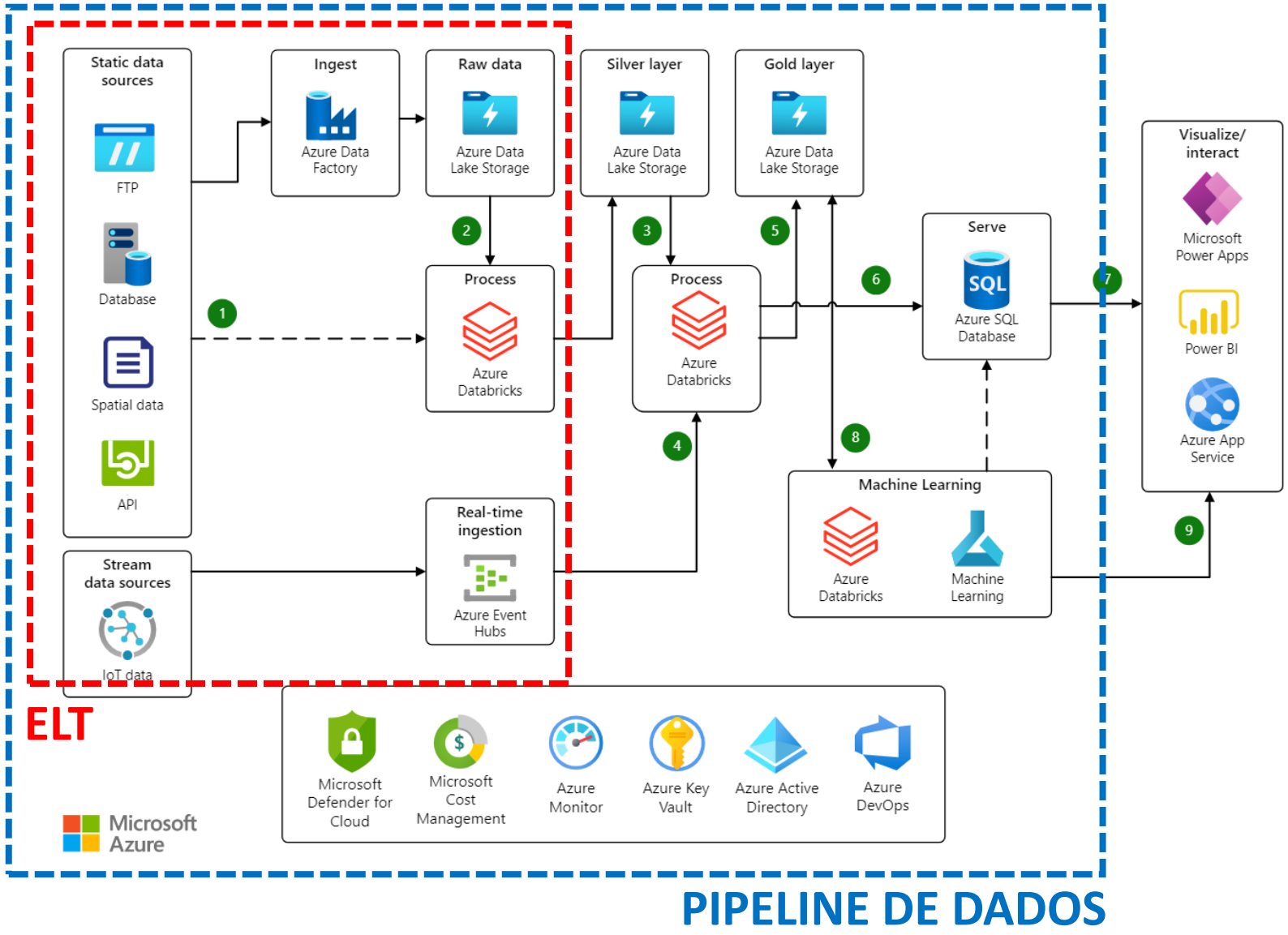
Pipeline de Dados
ETL X ELT



Fonte da imagem: https://www.linkedin.com/pulse/era-do-lakehouse-uma-evolu%C3%A7%C3%A3o-das-estruturas-de-dados-david-alpa-kmlnf/

Pipeline de Dados

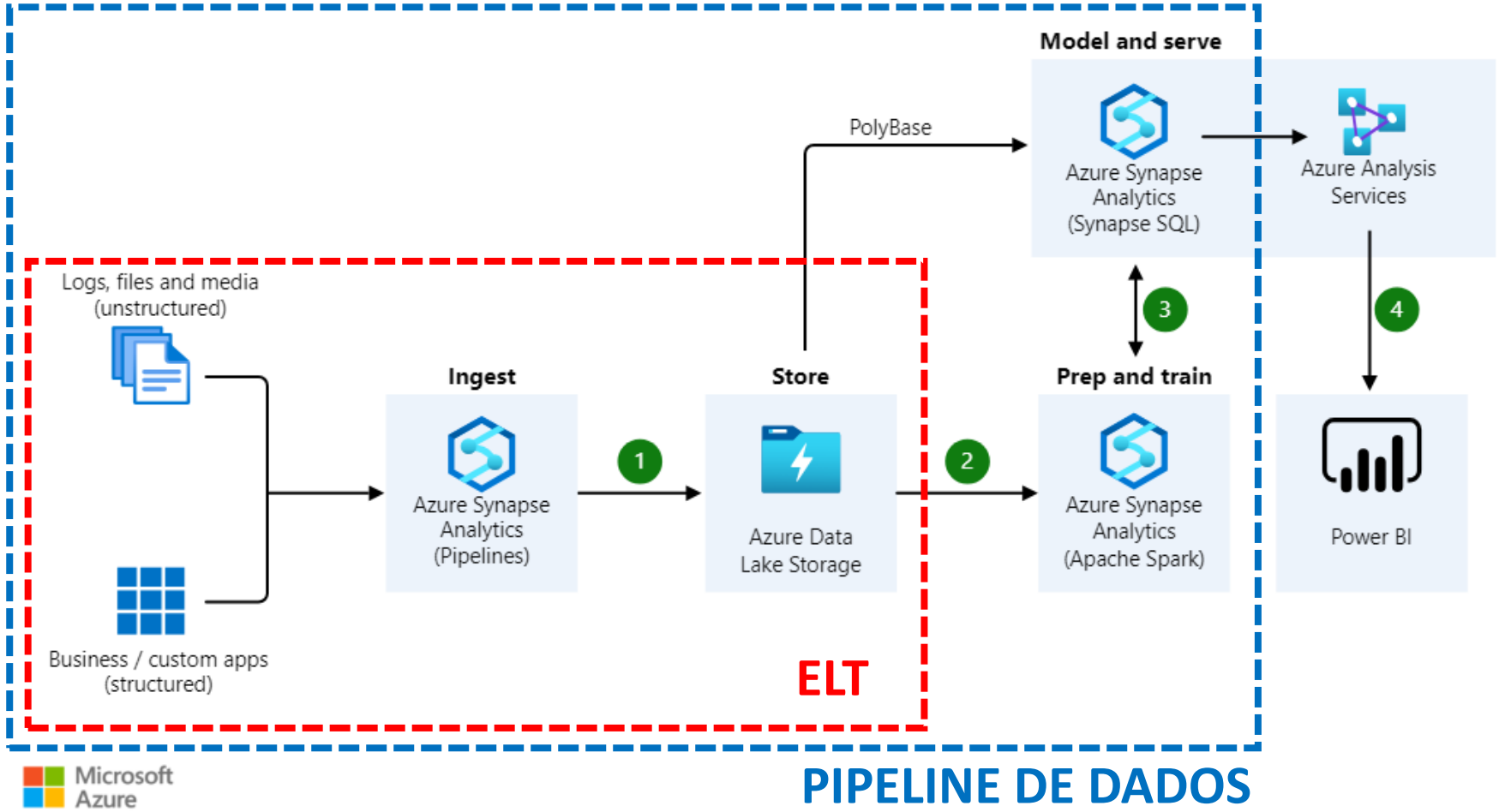
EXEMPLOS DE PIPELINE DE DADOS E ETL



Fonte da imagem: https://learn.microsoft.com/pt-br/azure/architecture/example-scenario/analytics/sports-analytics-architecture-azure

Pipeline de Dados

EXEMPLOS DE PIPELINE DE DADOS E ETL



Fonte da imagem: https://learn.microsoft.com/pt-br/azure/architecture/example-scenario/analytics/sports-analytics-architecture-azure

Atividade Prática

EXTRAÇÃO DE DADOS DE UM BANCO DE DADOS SQL E GRAVAR EM CSV

1. Subir um banco de dados SQL Server via Docker, conectar via SSMS e criar as tabelas e dados do ambiente relacional.

2. Criar um Azure Data Lake Storage Gen2 através do Terraform e Azure CLI
https://github.com/jlsilva01/adls-azure/

3. Criar um script em Python que leia os dados das tabelas do SQL e grave no formato CSV na camada landing-zone do Azure ADLS.

4. Executar os notebooks da aula passada no Databricks para levar o dado da landing-zone até a camada gold.

BONUS: Criar um banco de dados MongoDB via Docker e executar o passo 3 desta atividade.