

# **Análise de Dados com Base em Processamento Massivo em Paralelo**

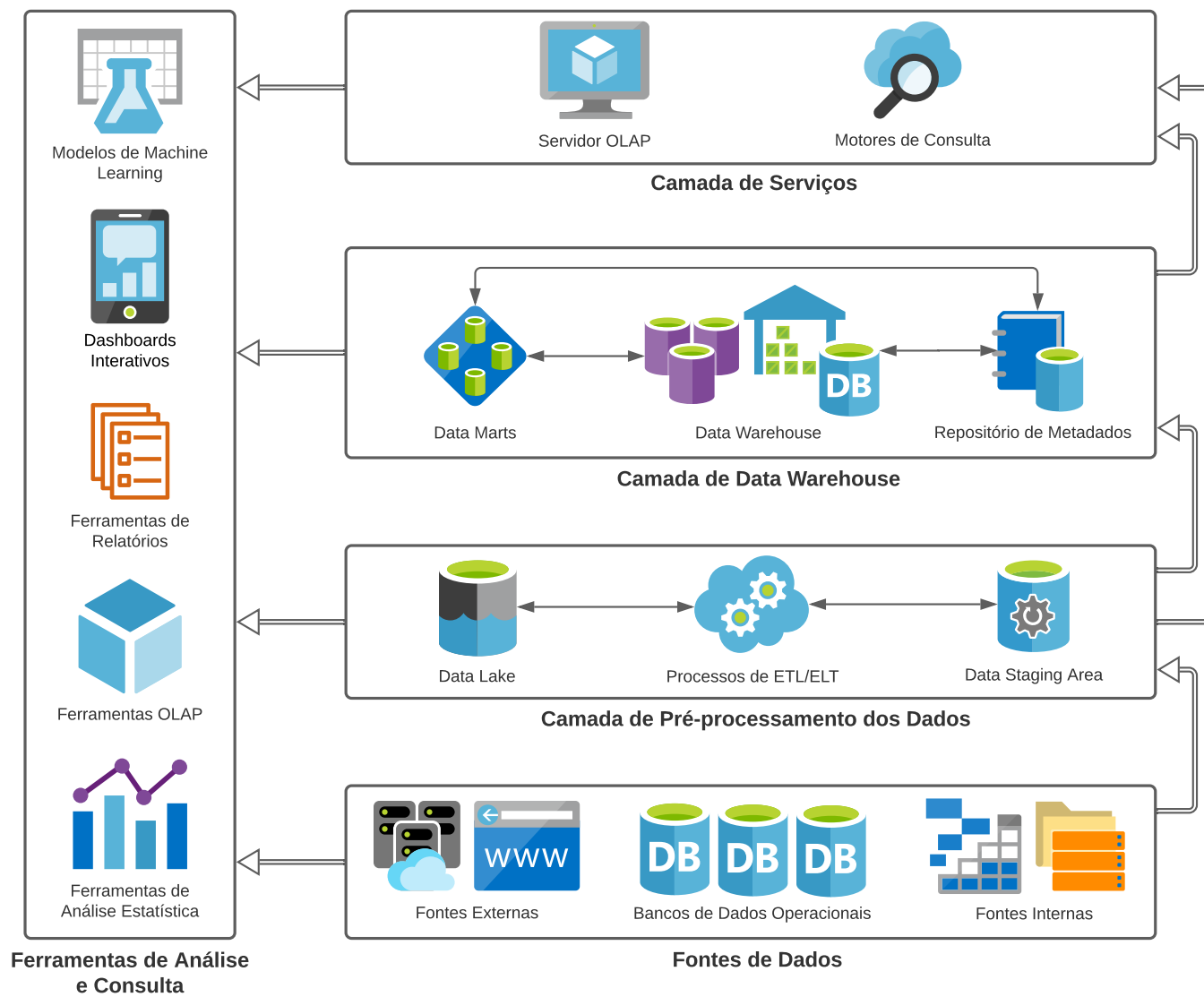
## **Aula 2: Arquitetura de Data Warehousing**

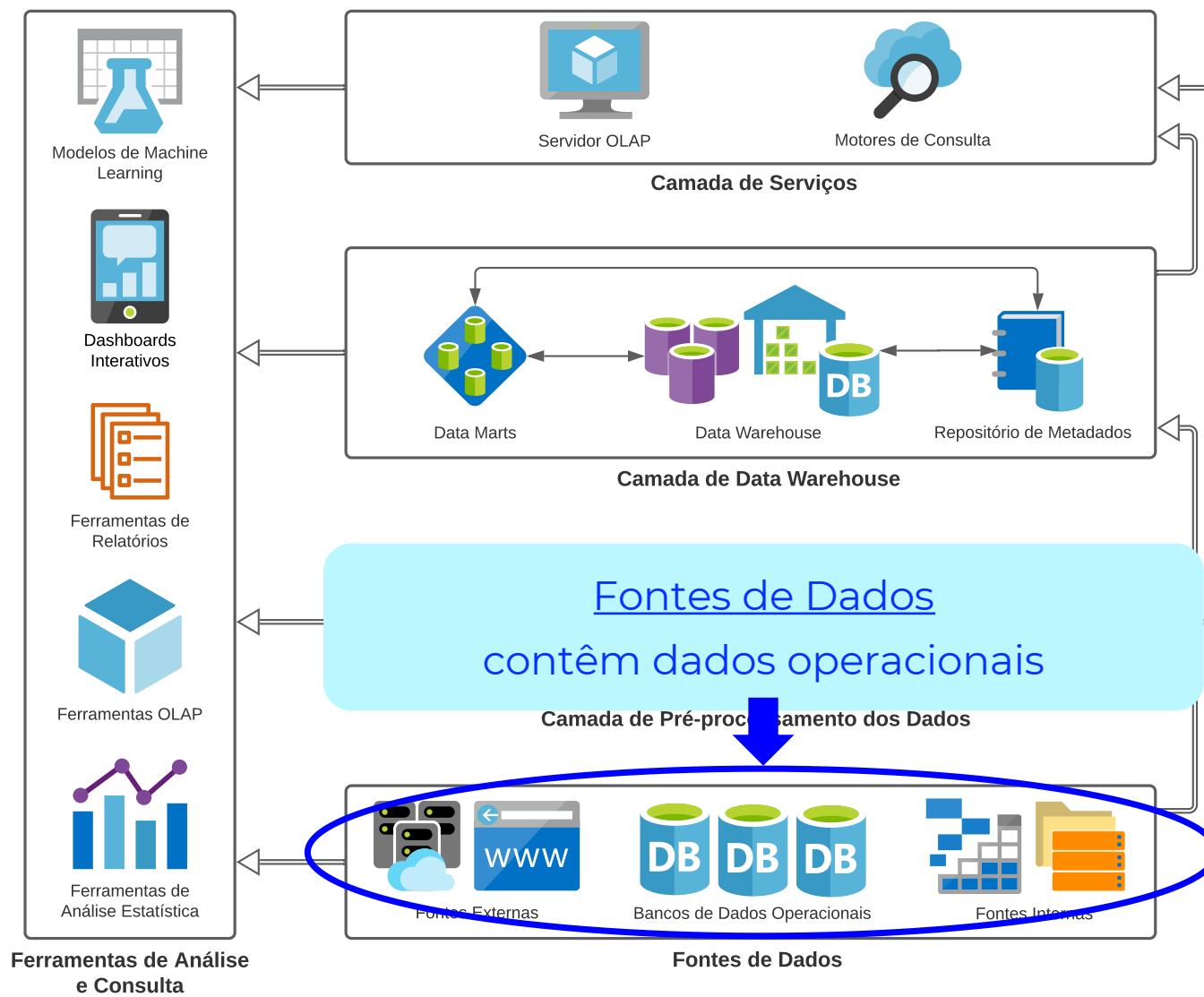
Cristina Dutra de Aguiar Ciferri  
ICMC/USP  
cdac@icmc.usp.br

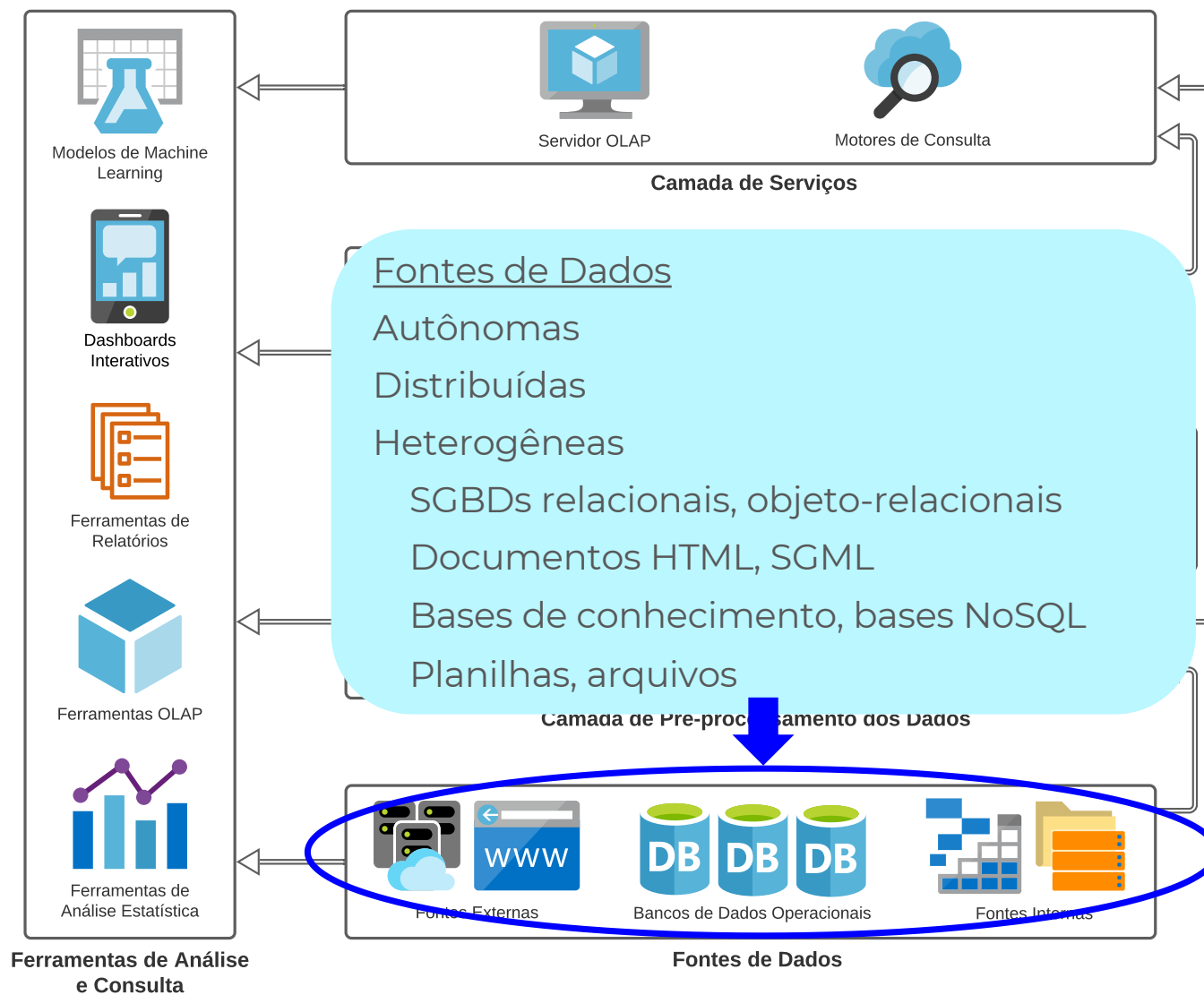


# Agenda

- Visão Geral
- Diferenças entre os Locais de Armazenamento
- Big Data
- Exemplos de Pipeline





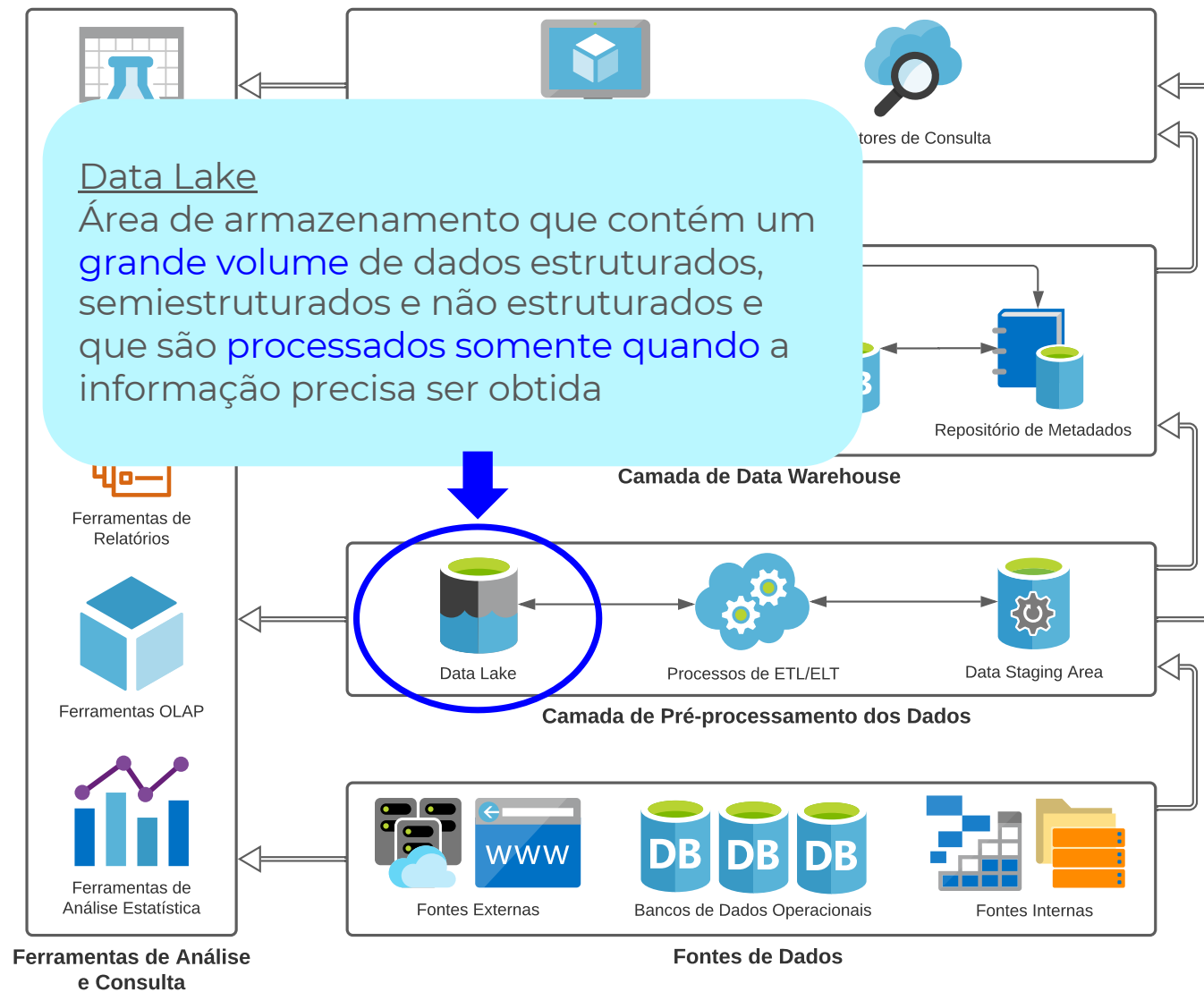




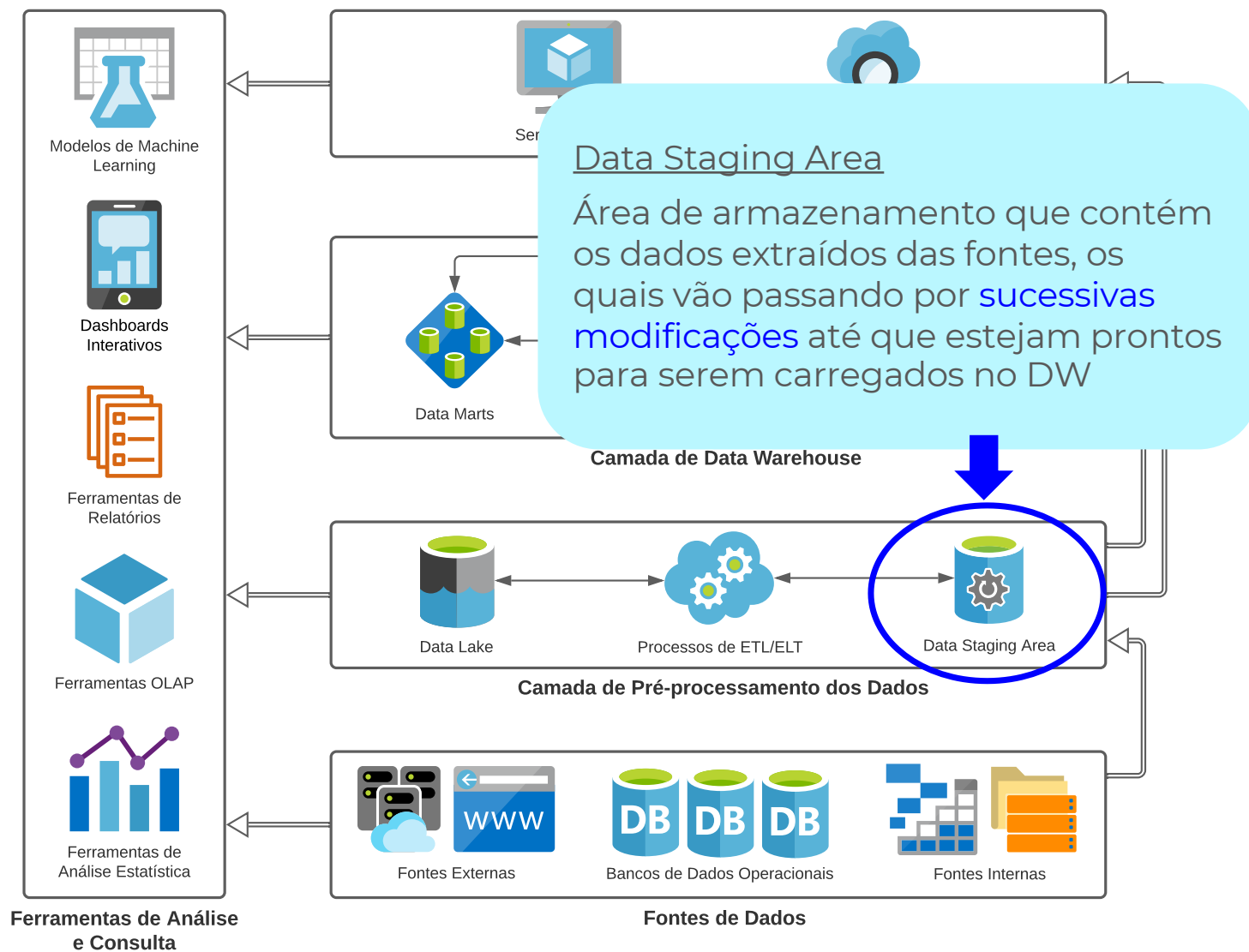


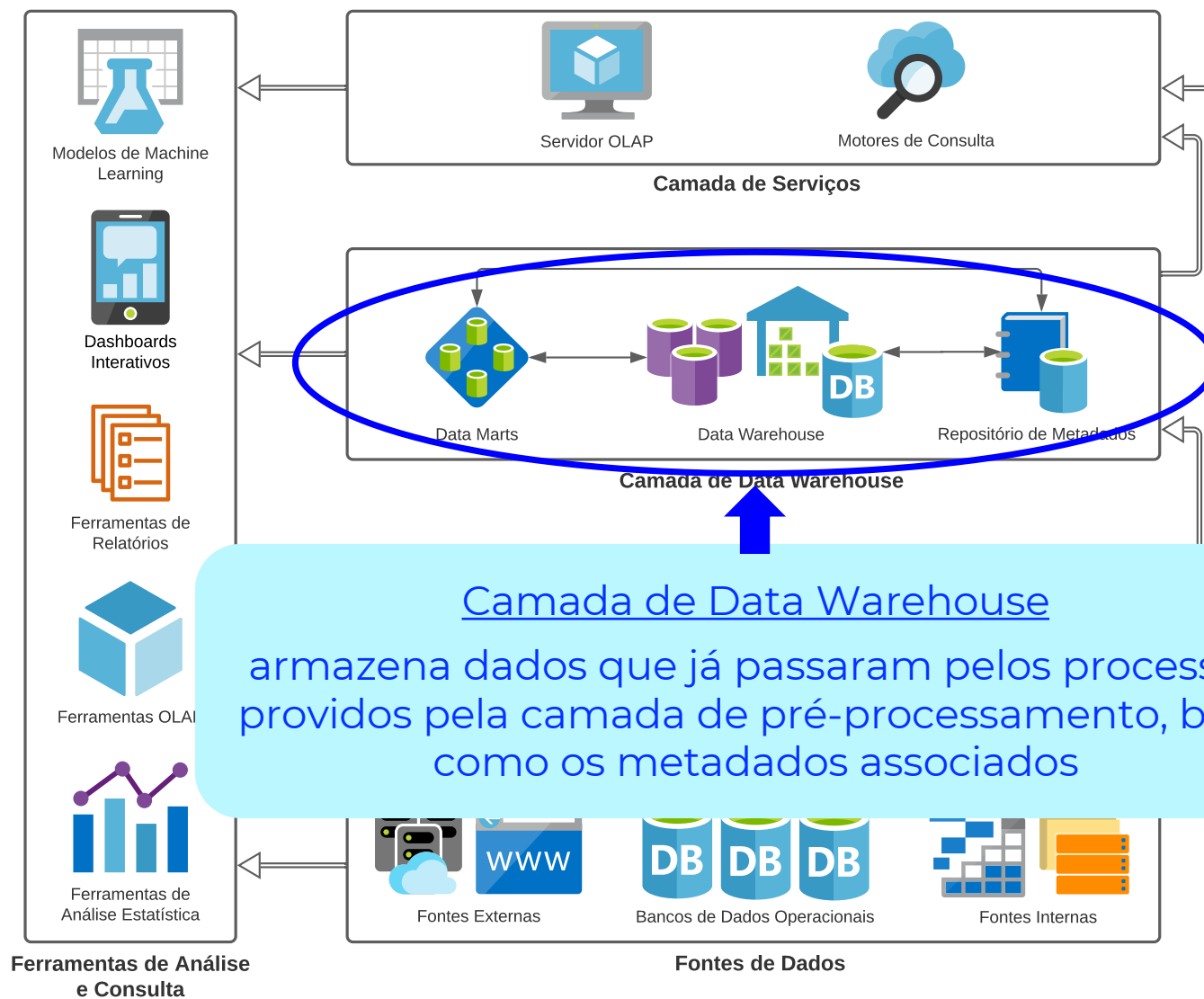
## Data Lake

Área de armazenamento que contém um **grande volume** de dados estruturados, semiestruturados e não estruturados e que são **processados somente quando** a informação precisa ser obtida









### Camada de Data Warehouse

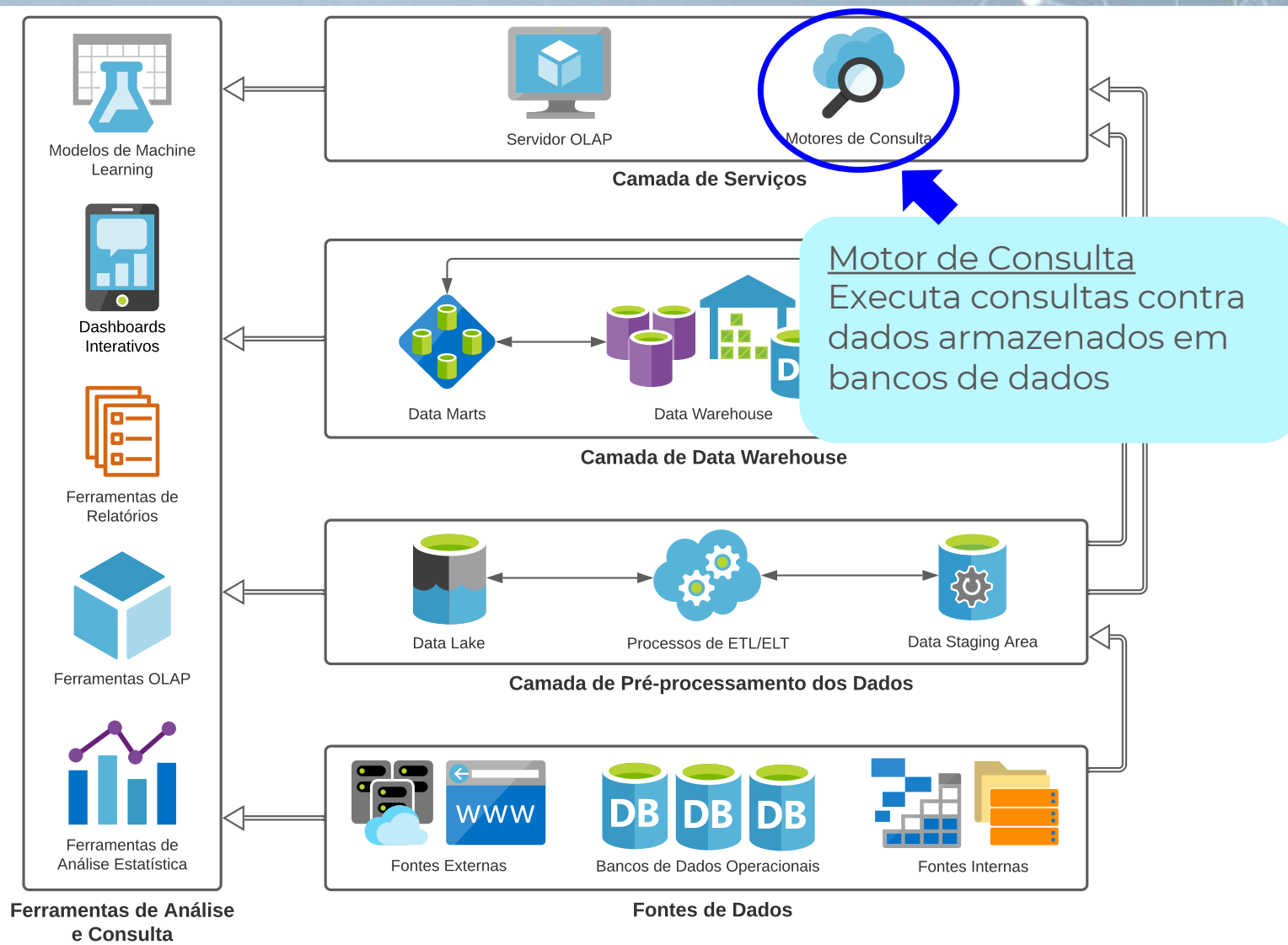
armazena dados que já passaram pelos processos  
providos pela camada de pré-processamento, bem  
como os metadados associados





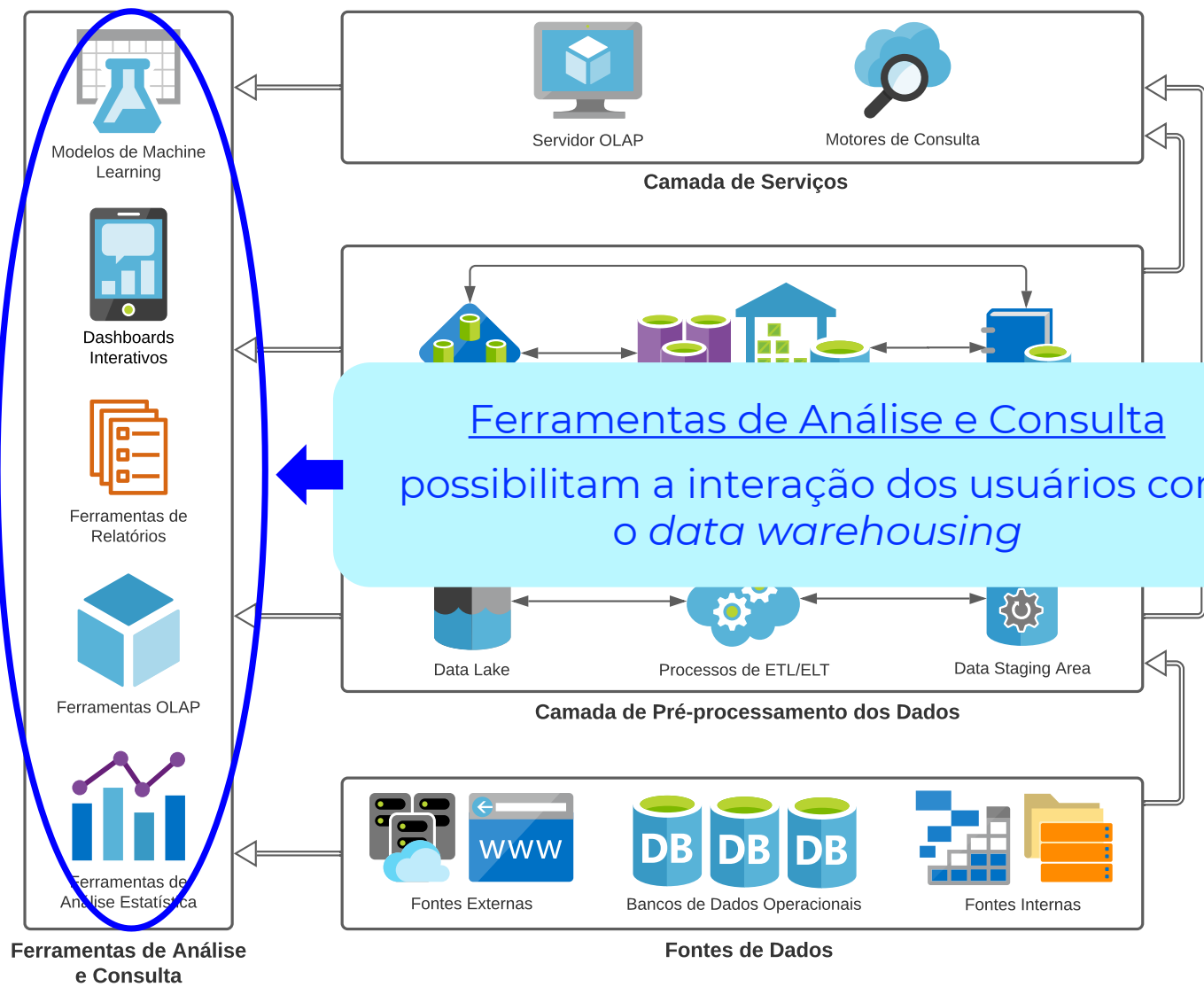


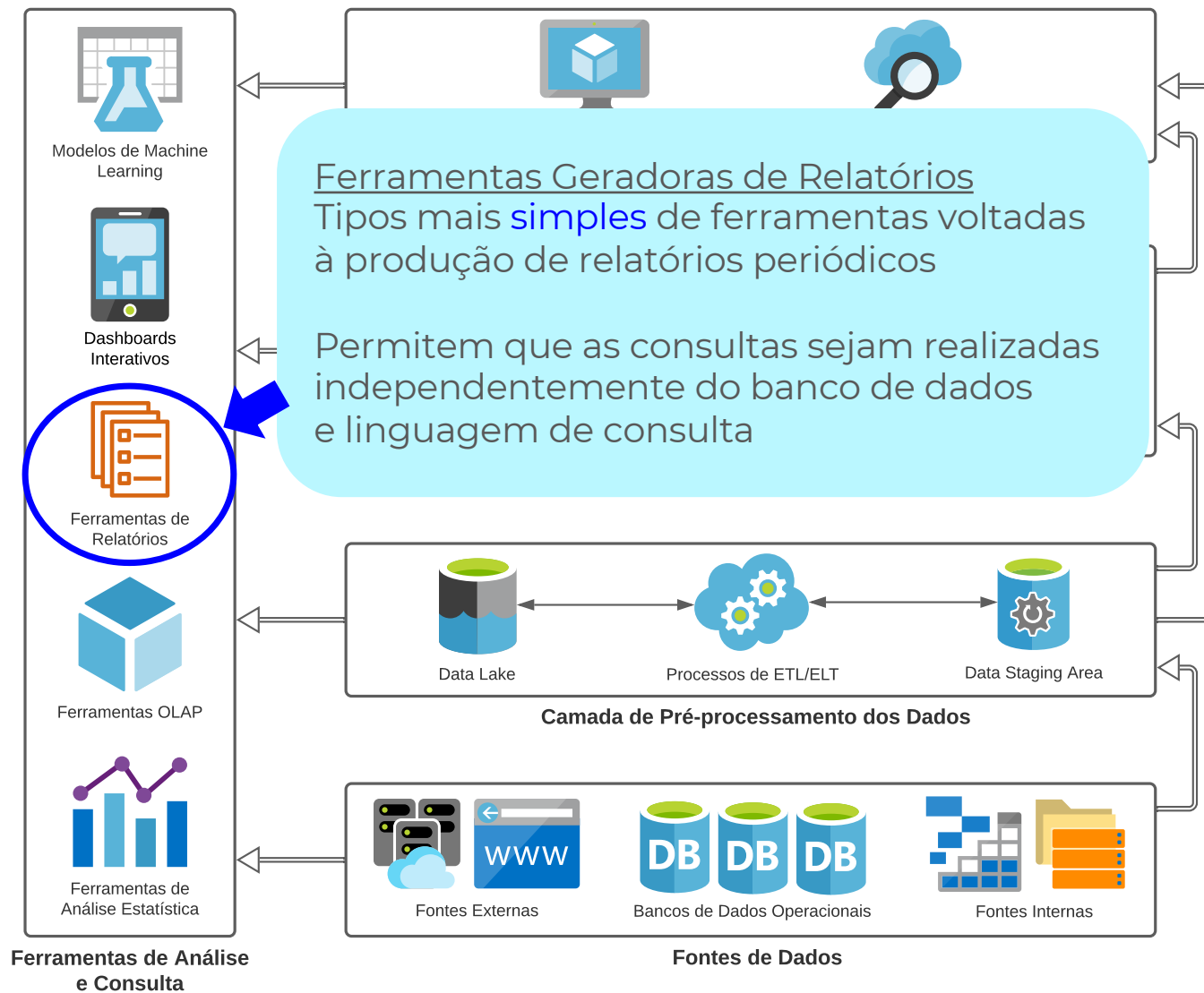




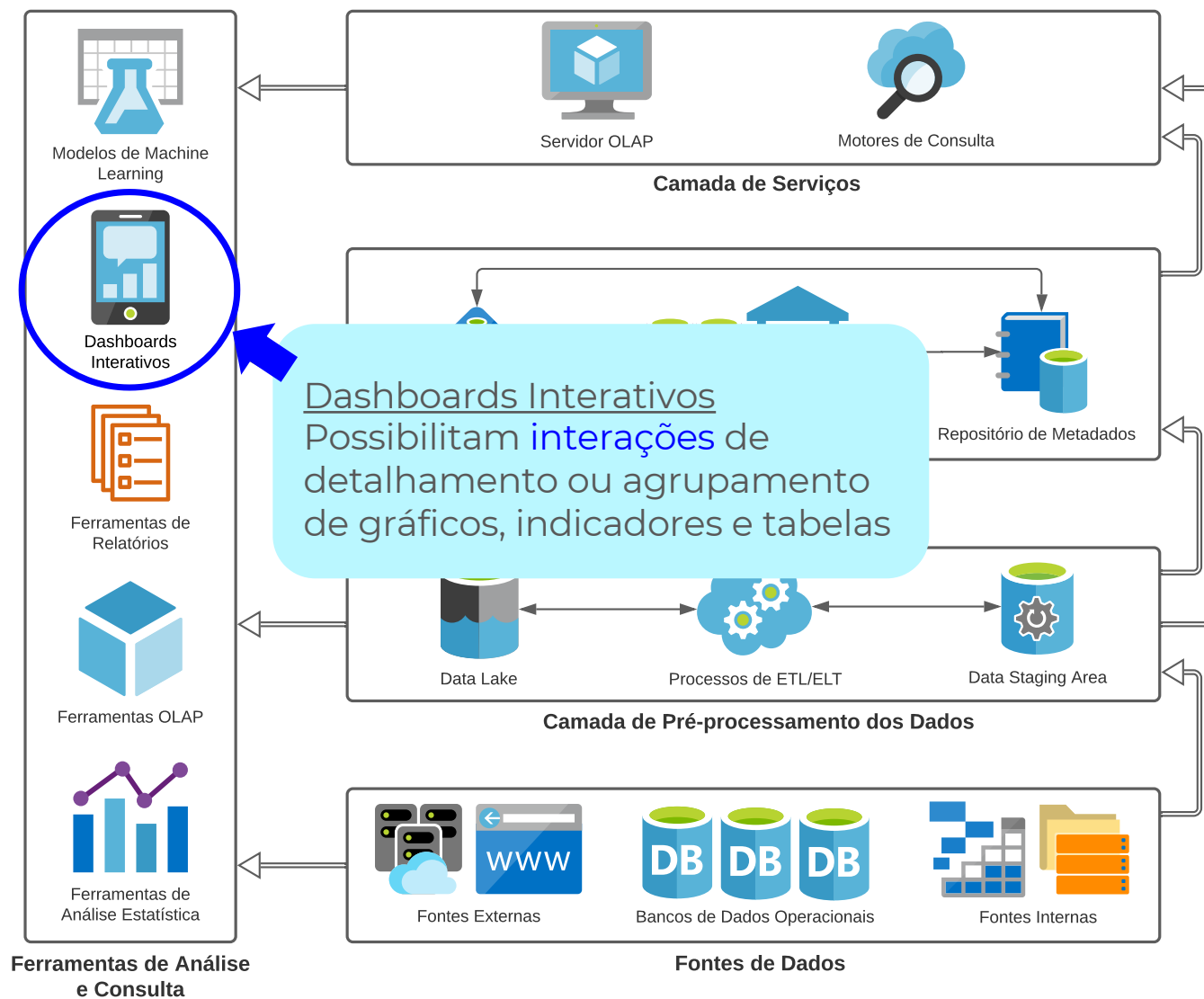


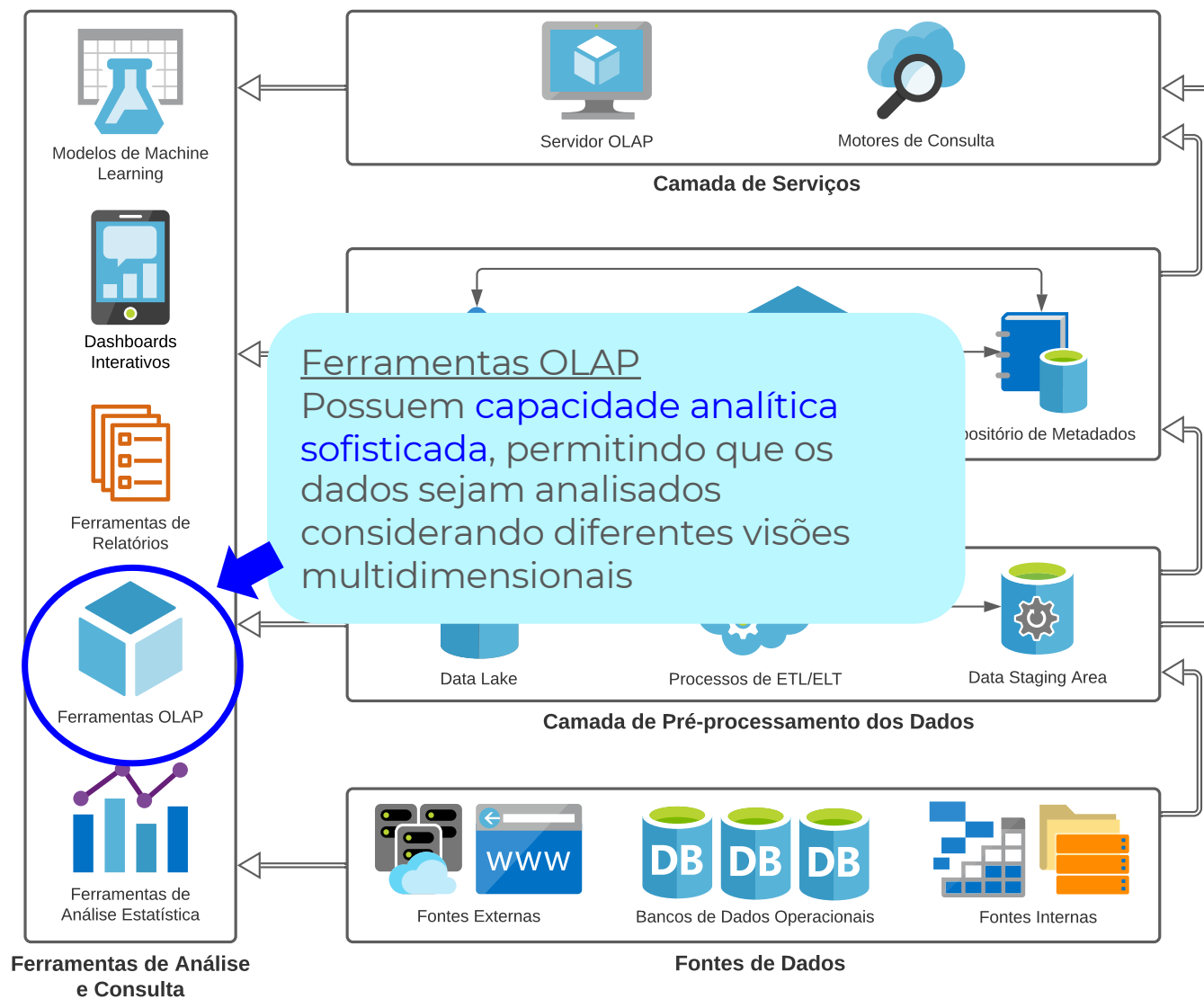


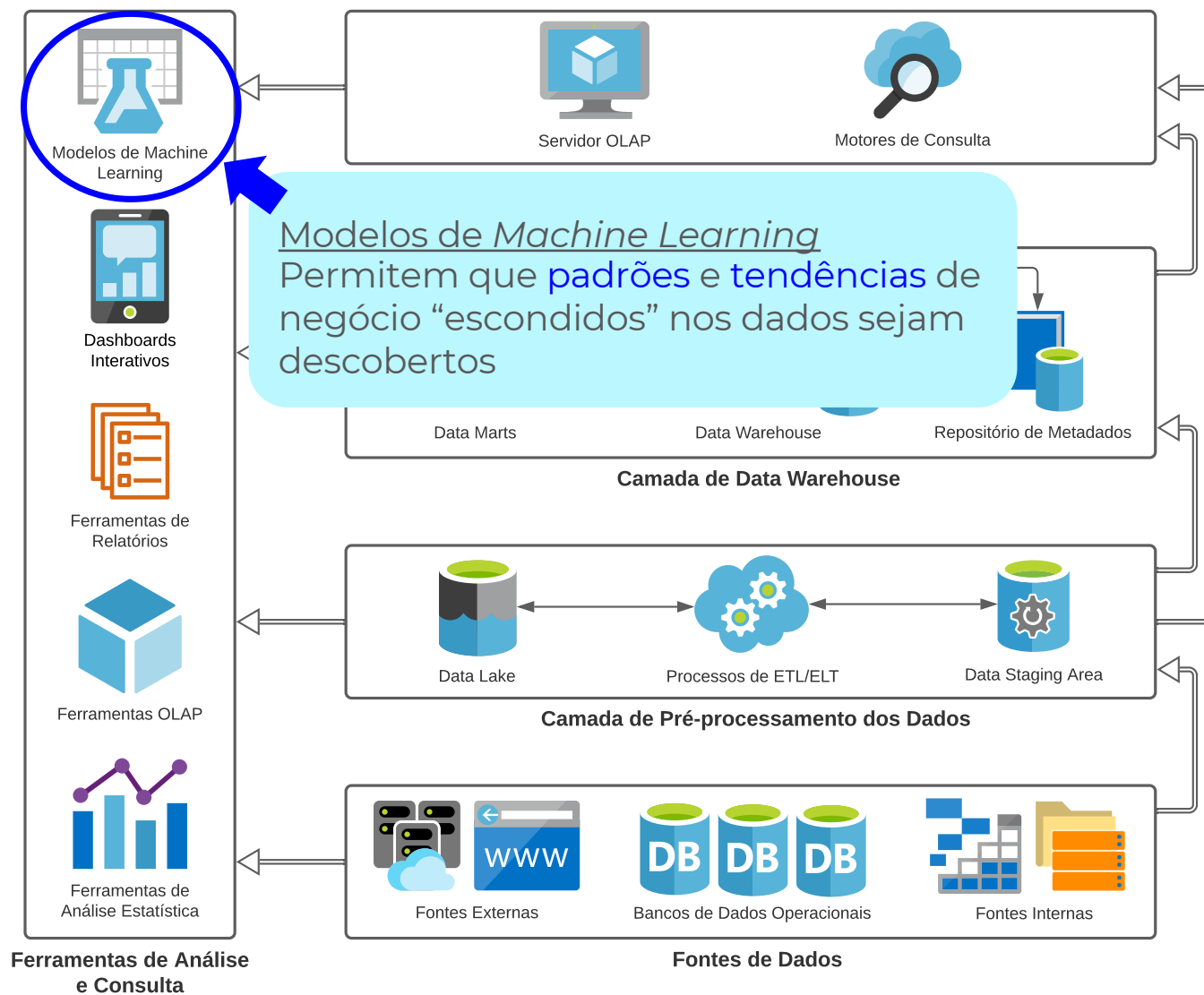


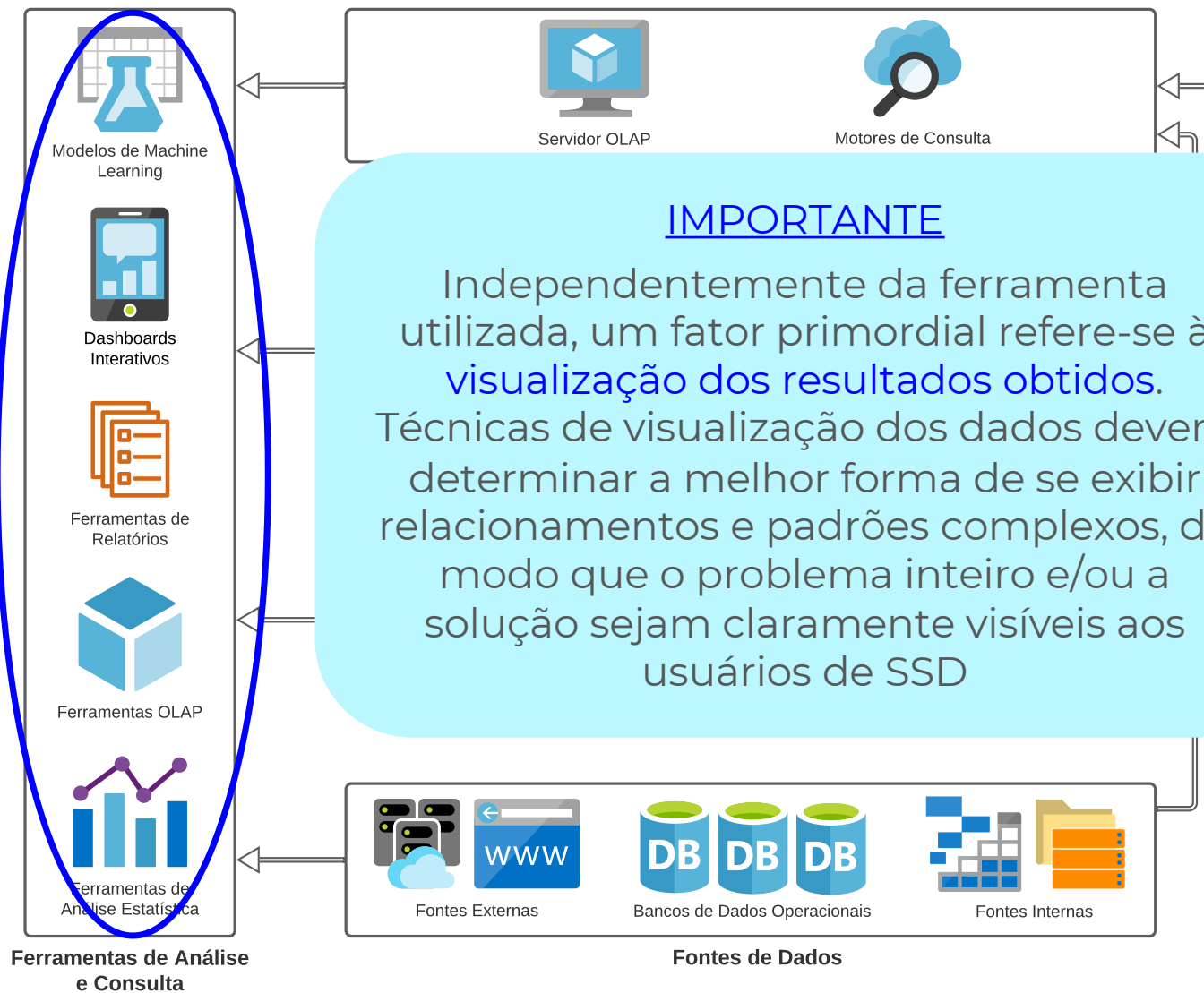








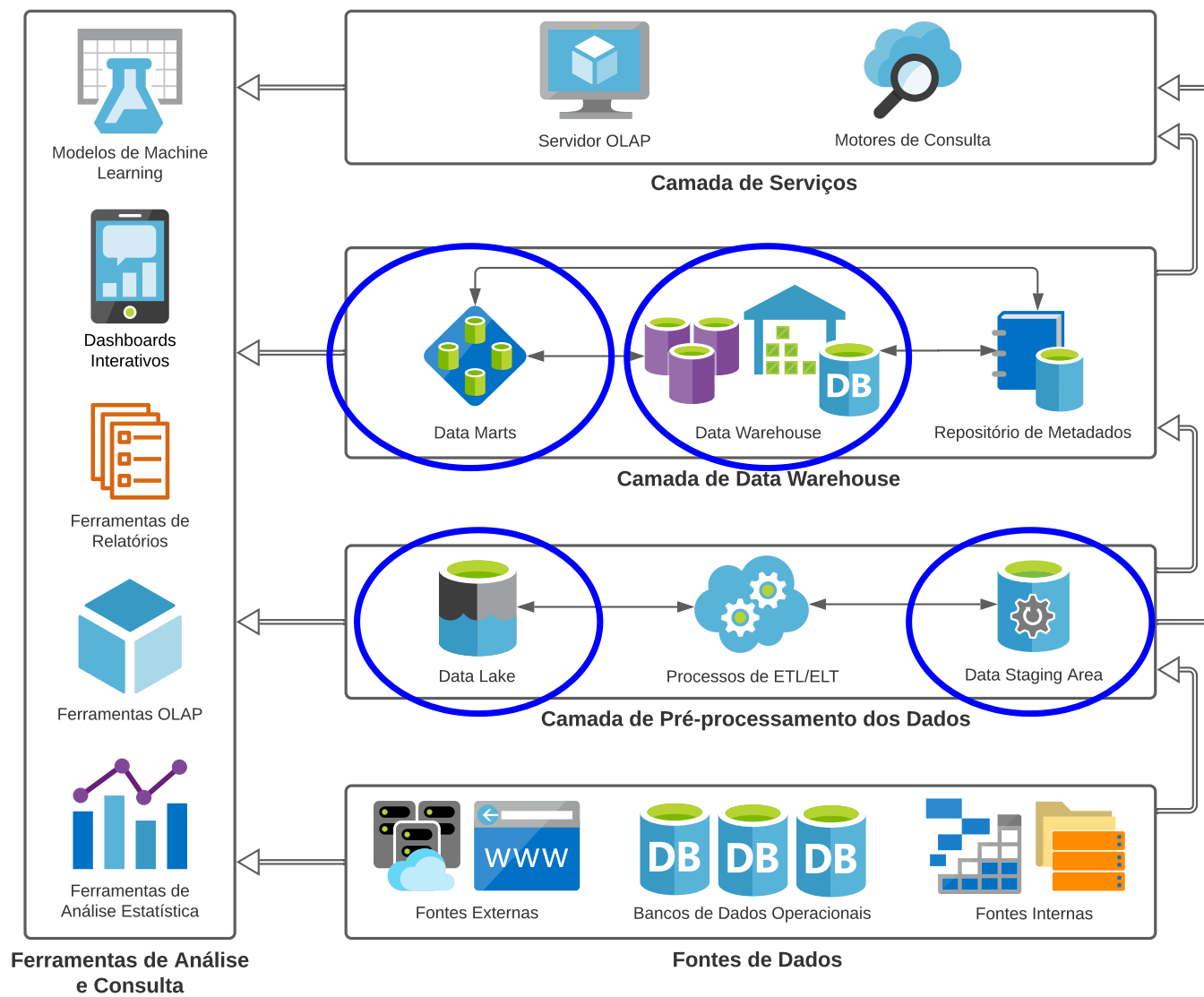


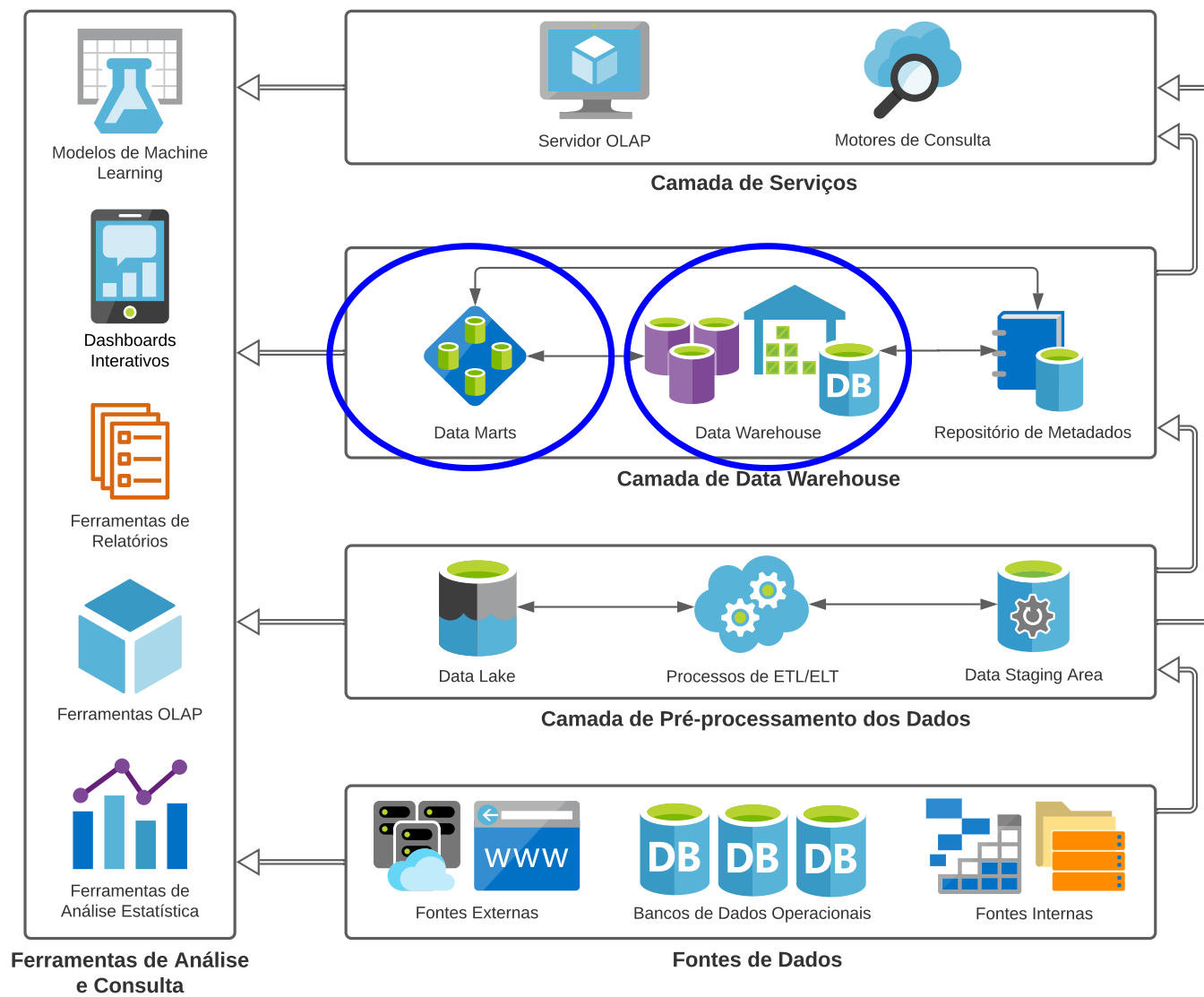


# Agenda

- Visão Geral
- Diferenças entre os Locais de Armazenamento
- Big Data
- Exemplos de Pipeline







# Diferenças entre Data Warehouse e Data Mart

- Data Mart
  - Consiste de um DW com **escopo limitado**
  - Armazena dados que possuem as **mesmas características** dos dados do DW
- Política de construção evolucionária de um DW corporativo
  - Processo de construção de um DW corporativo é longo, complexo e demanda alto investimento financeiro
  - Construção paulatina de vários **data marts independentes**, cada um atendendo um assunto de interesse específico

# Exemplo: BI Solutions



Demanda: analisar gastos  
com salários de funcionários

Data Mart 1

Foco: [salários](#) e  
[quantidadeLançamentos](#)

Perspectivas: funcionário

cargo

filial

data

Demanda: analisar gastos  
com material de consumo

Data Mart 2

Foco: [gastosConsumo](#)

Perspectivas: material

filial

data

Demanda: analisar gastos  
com equipamento  
de infraestrutura

Data Mart 3

Foco: [gastosInfra](#)

Perspectivas: equipamento

filal

data

fornecedor

# Exemplo: BI Solutions



Demanda: analisar receitas  
relativas aos cursos de treinamento  
dos produtos vendidos

Demanda: analisar receitas  
relativas às vendas dos  
produtos da empresa

## Data Mart 4

Foco: receitaVendas

Perspectivas: produto  
cliente  
filial  
data

## Data Mart 5

Foco: receitaTreinamento

Perspectivas: funcionário  
cargo  
data  
cliente  
produto

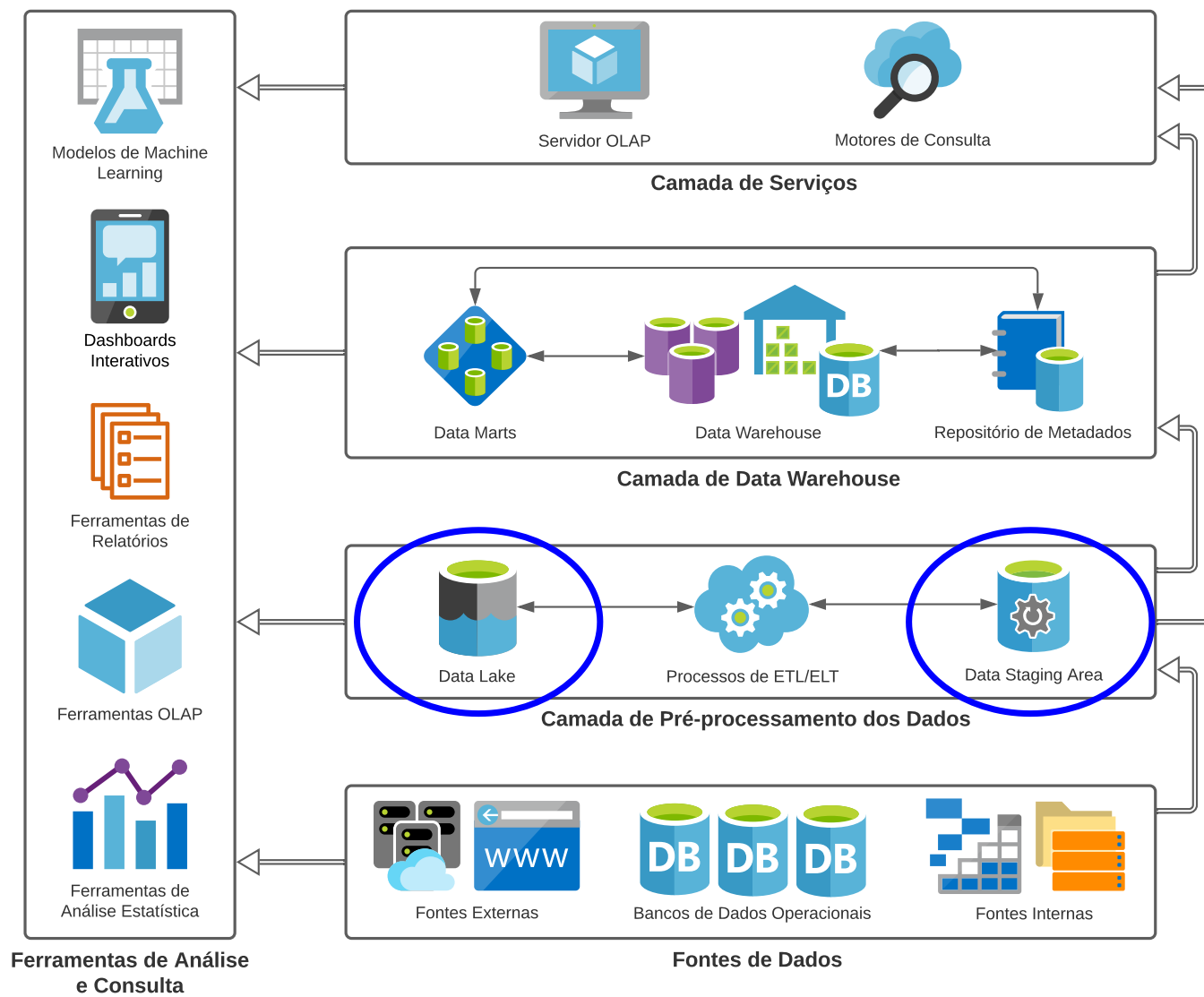
# Uso de Data Marts Independentes

- Vantagens

- Reduz gastos financeiros iniciais, desde que exige recursos monetários inferiores aos despendidos com a construção de um DW corporativo
- Possibilita que usuários de SSD reconheçam o valor e a potencialidade da solução de *data warehousing* em um período menor de tempo

- Desvantagens

- Pode conduzir a diferentes problemas caso um modelo de negócio completo não seja bem especificado e desenvolvido de acordo
- Cada *data mart* independente pode tornar-se autônomo, heterogêneo e distribuído



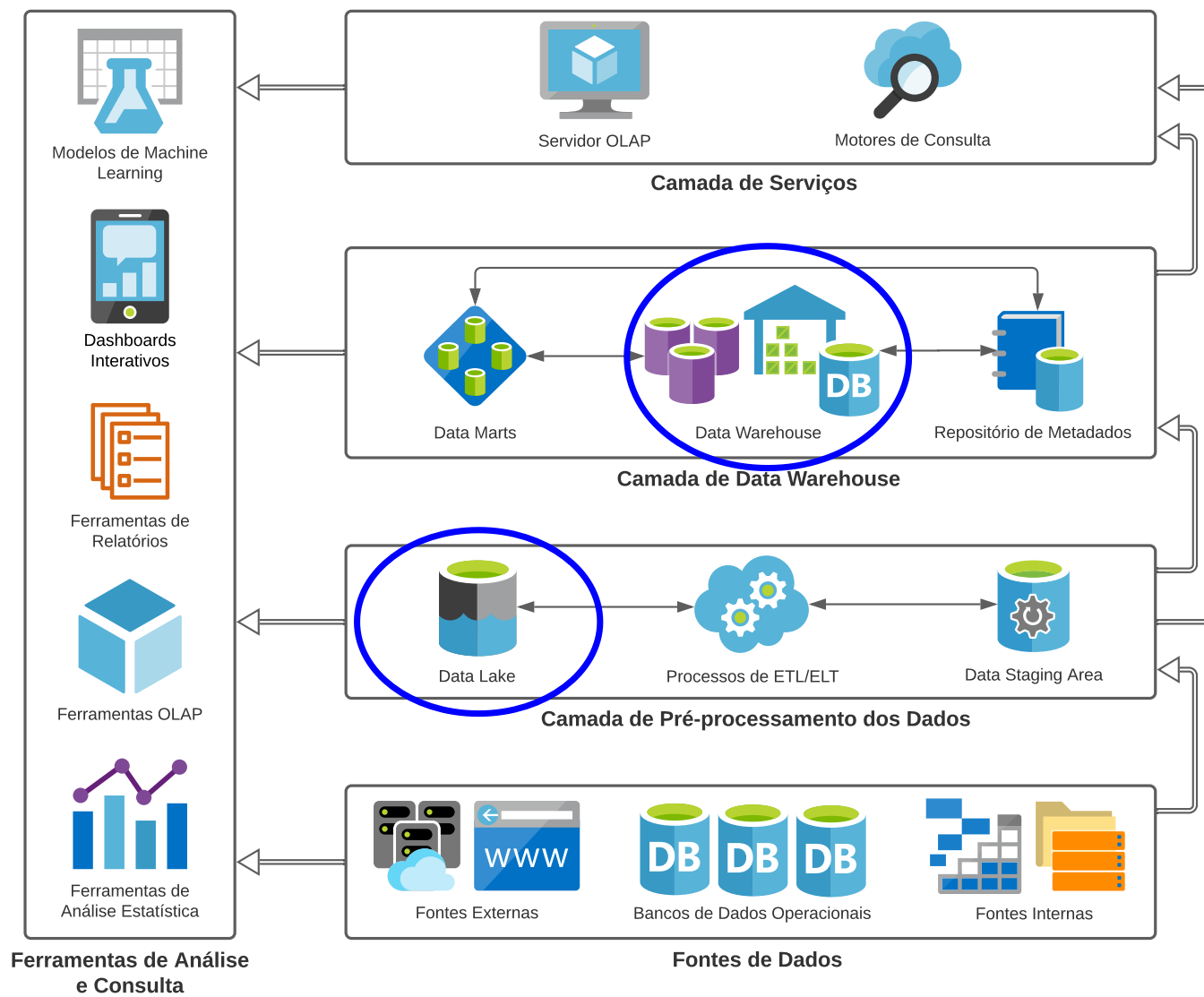
# Diferenças entre Data Staging Area e Data Lake

	Data Staging Area	Data Lake
Dados Armazenados	dados que sofrem modificações sucessivas	dados no formato nativo ( <i>raw data</i> ), incluindo dados estruturados, semiestruturados e não estruturados
Processamento dos Dados	dados prontos para serem carregados no DW	dados processados somente quando a informação precisa ser obtida



# Diferenças entre Data Staging Area e Data Lake

	Data Staging Area	Data Lake
Fluxo de Dados	<i>data staging area</i> → DW	<i>data lake</i> → DW <i>data lake</i> → ferramentas de análise e consulta
Decorrência Histórica	processo ETL	processo ELT



# Diferenças entre Data Warehouse e Data Lake

	Data Warehouse	Data Lake
Característica dos Dados	consolidados, organizados e estruturados	estruturados, semiestruturados e não estruturados
Formato dos Dados	esquema estruturado (formato bem definido)	formato nativo (diferentes formatos)
ETL/ELT	dados pré-processados antes de serem carregados	dados extraídos e carregados, sem sofrer transformações

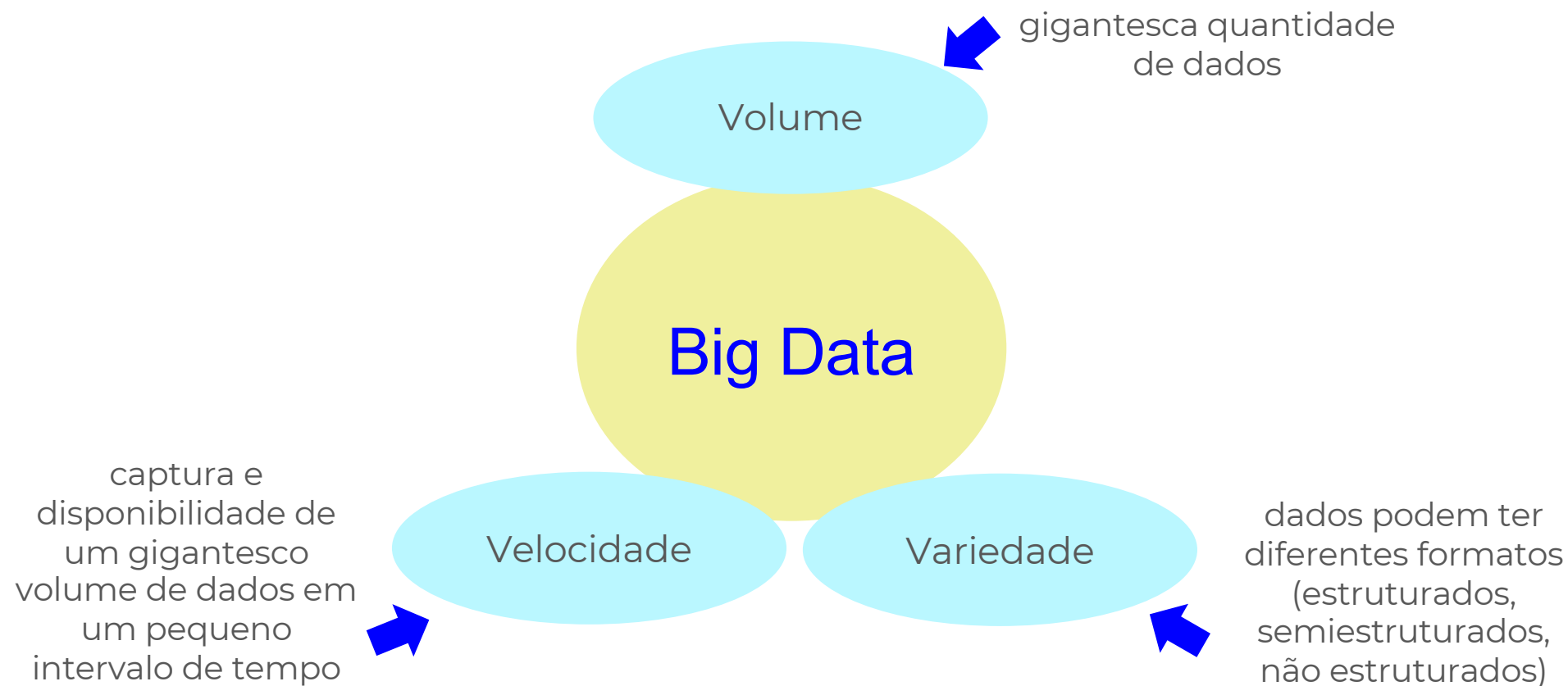
# Diferenças entre Data Warehouse e Data Lake

	Data Warehouse	Data Lake
<b>Tipos de Consulta</b>	OLAP	variado
<b>Latência para Disponibilizar os Dados</b>	alta	baixa
<b>Custo de Geração dos Dados</b>	maior	menor
<b>Custo de Análise dos Dados</b>	menor	maior

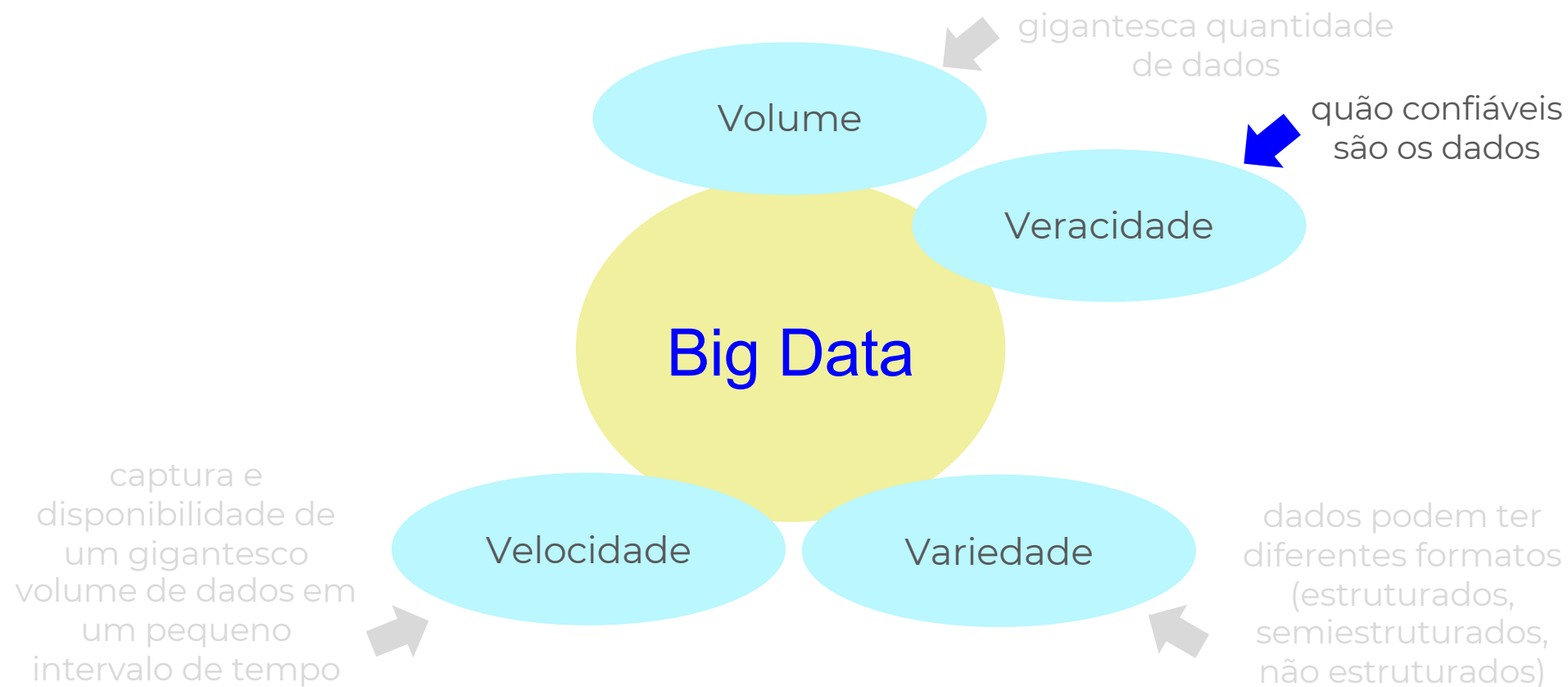
# Agenda

- Visão Geral
- Diferenças entre os Locais de Armazenamento
- Big Data
- Exemplos de Pipeline

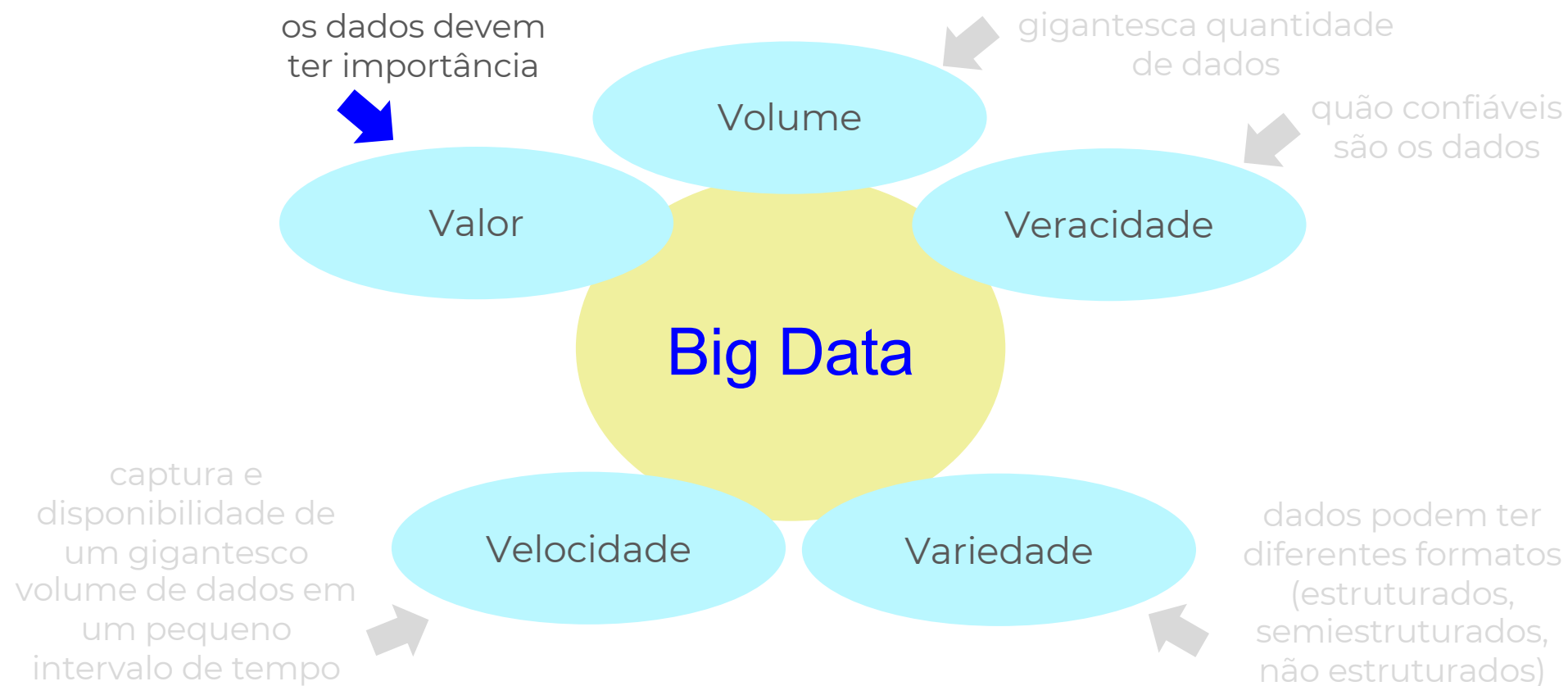
# Modelo de 3Vs



# Modelo de 4Vs

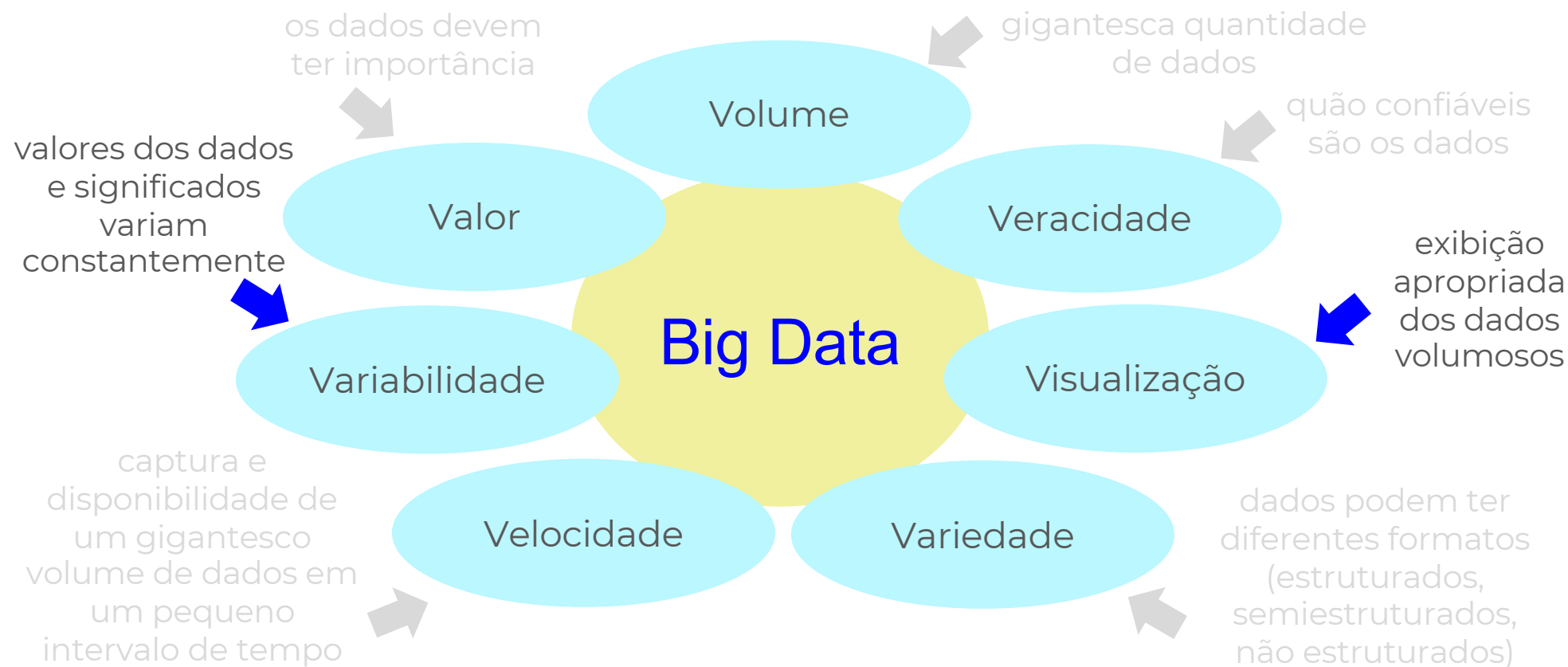


# Modelo de 5Vs





# Modelo de 7Vs



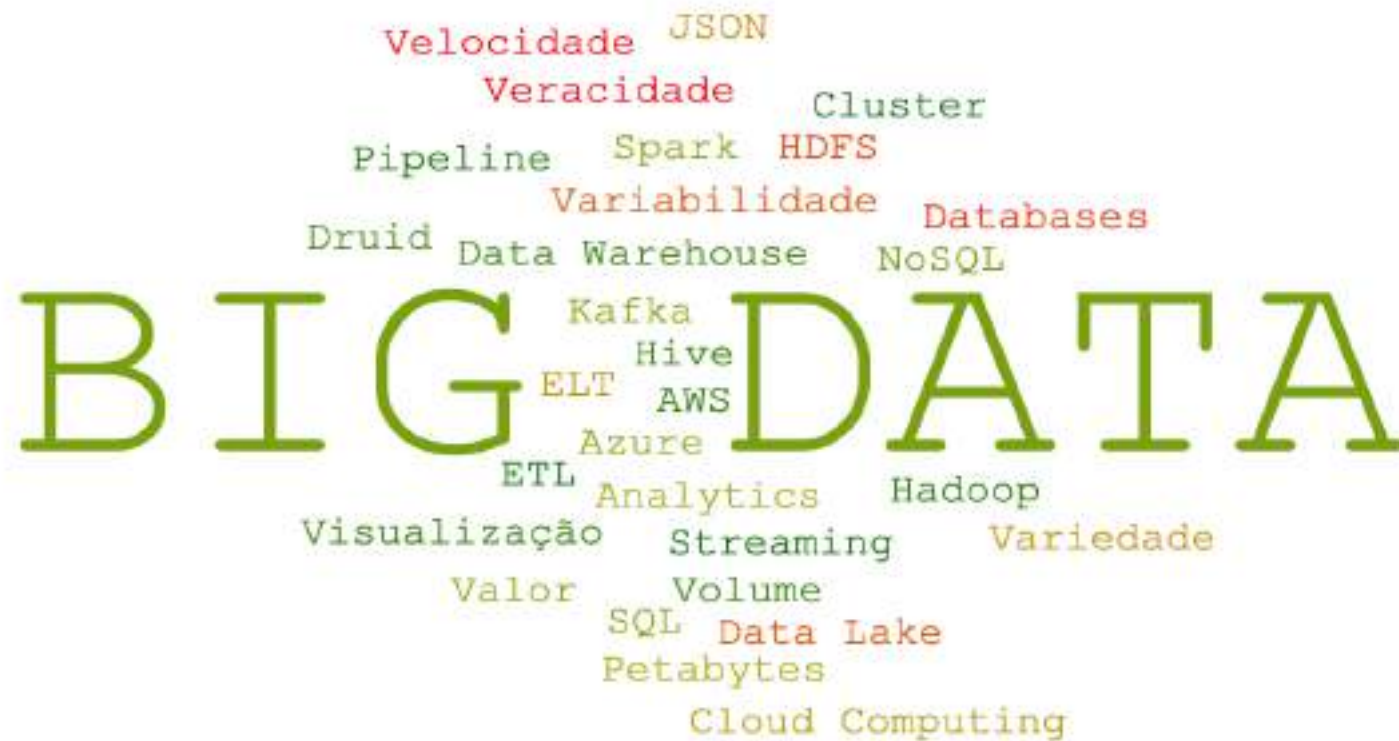
# Desafios

- Uso de **ambientes computacionais** com grande capacidade de armazenamento e processamento
  - *Clusters* de computadores
  - Computação em nuvem (*cloud computing*)
- Uso de **frameworks de processamento paralelo e distribuído** para simplificar a interação com os ambientes computacionais
  - Apache Hadoop
  - Apache Spark

# Desafios

- Uso de **sistemas de arquivos distribuídos** para prover suporte para o armazenamento de grandes quantidades de dados
  - HDFS (*Hadoop Distributed File System*)
- Uso de **bases de dados NoSQL (*Not only SQL*)** para introduzir flexibilidade no armazenamento de diferentes tipos de dados
  - Não estruturados
  - Semiestruturados
  - Estruturados

# Nuvem de Conceitos e Tecnologias



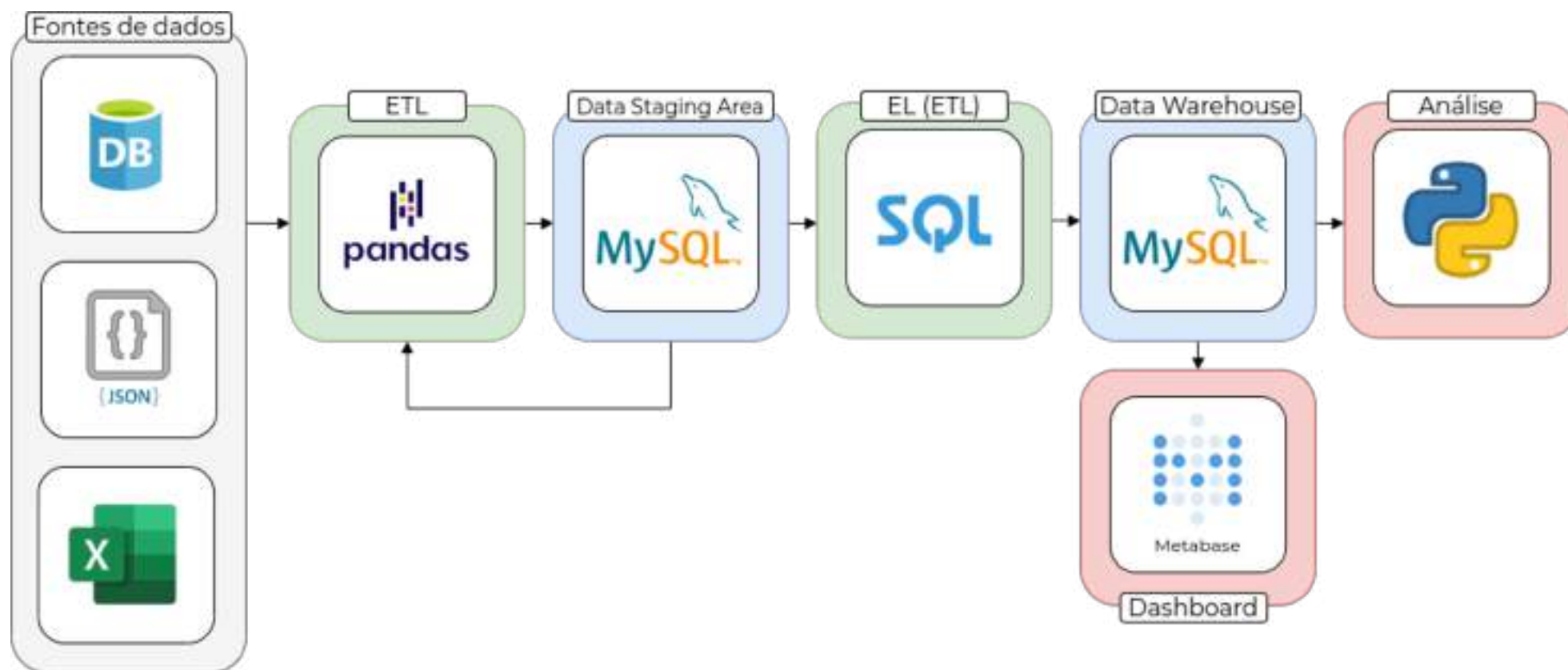
# Agenda

- Visão Geral
- Diferenças entre os Locais de Armazenamento
- Big Data
- Exemplos de Pipeline

# Exemplos de Pipeline

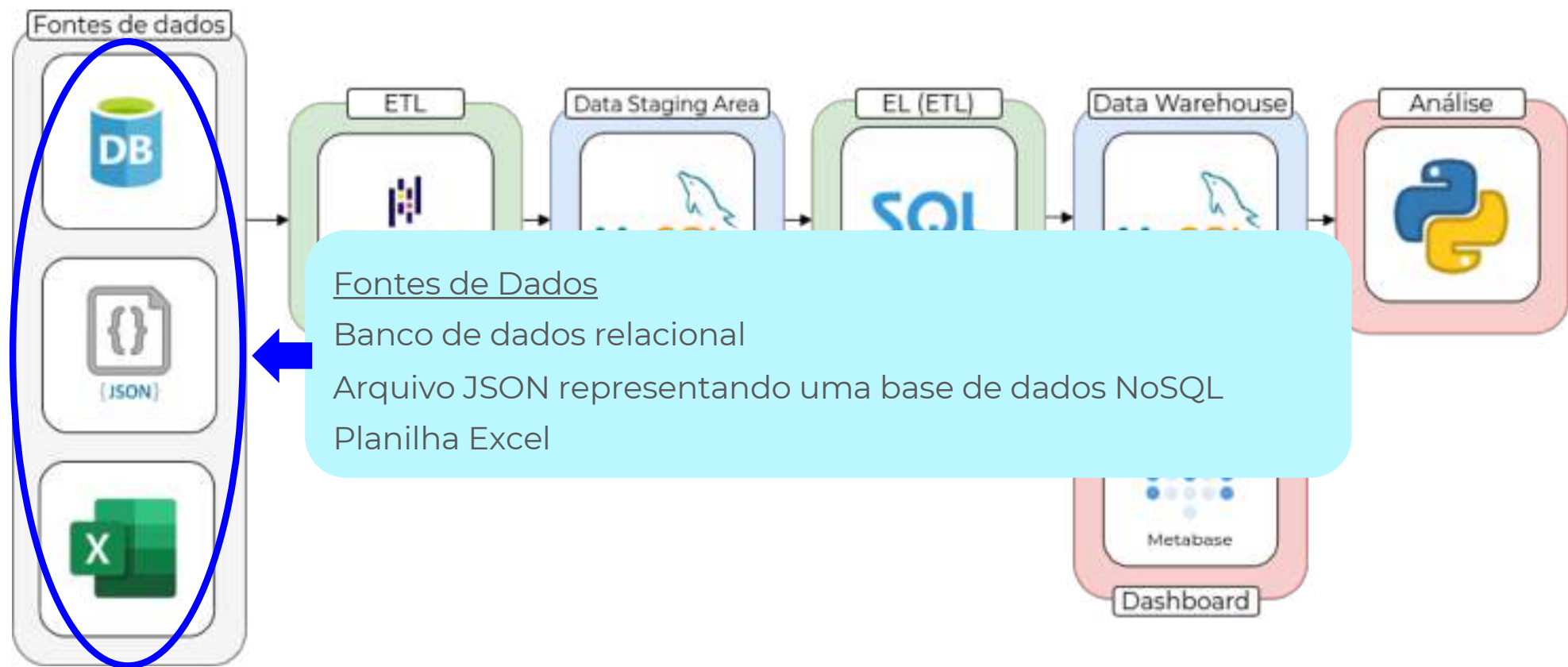
- Volumes de Dados Tradicionais
- Big Data
- Data Streaming

# Processamento de Dados em Lote



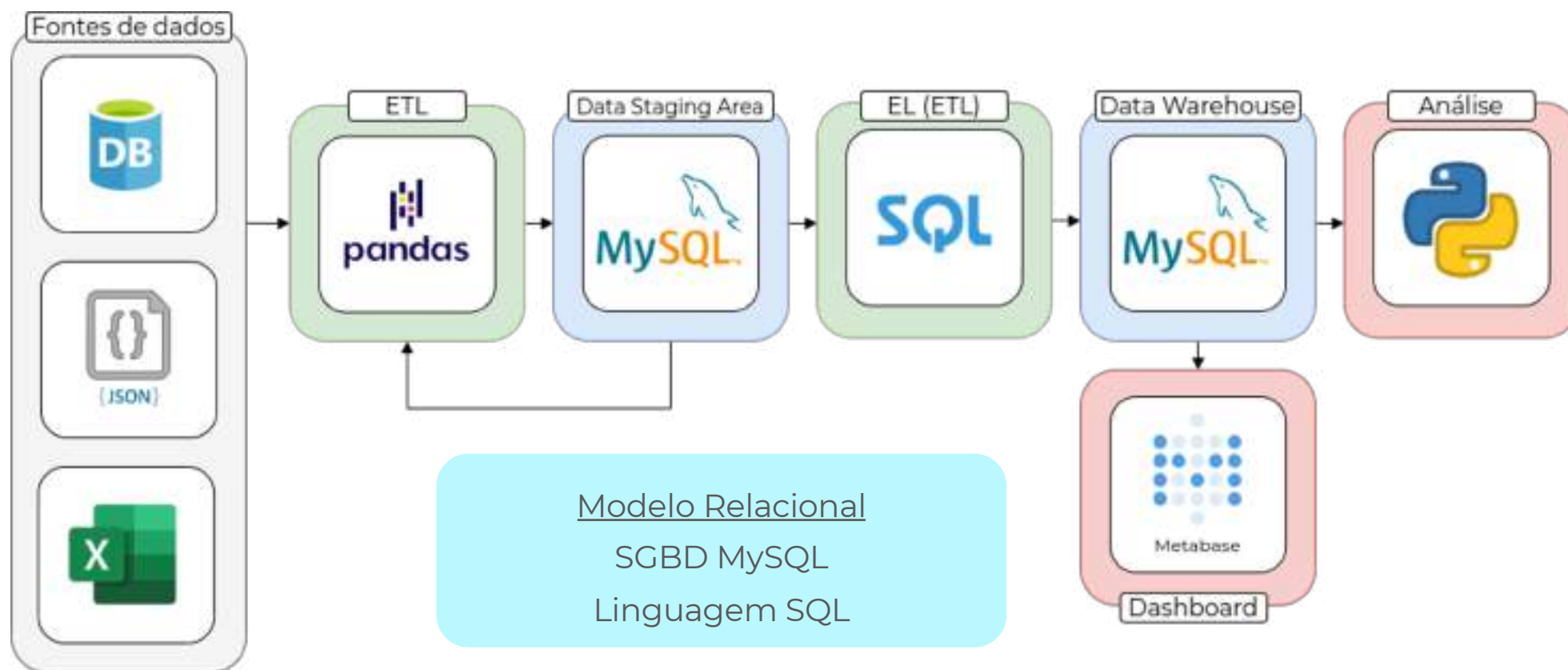


# Processamento de Dados em Lote

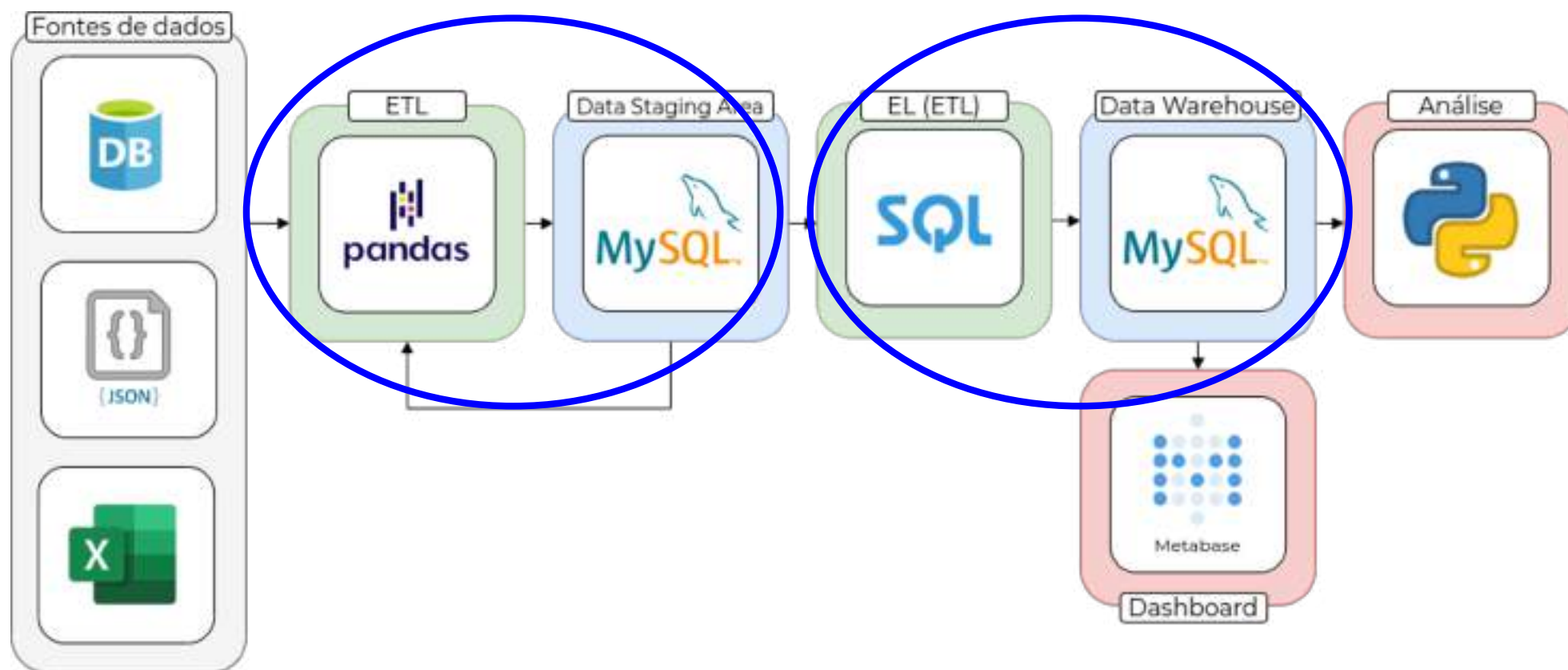




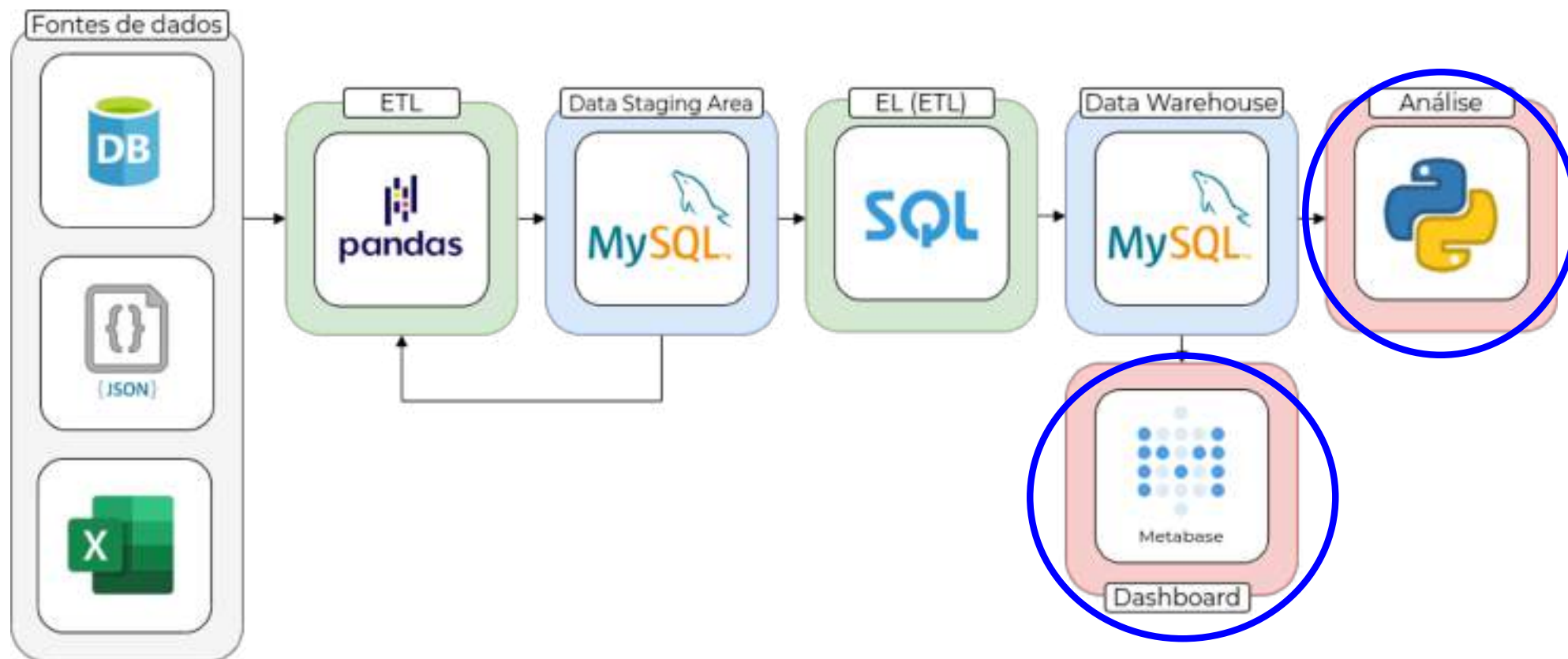
# Processamento de Dados em Lote



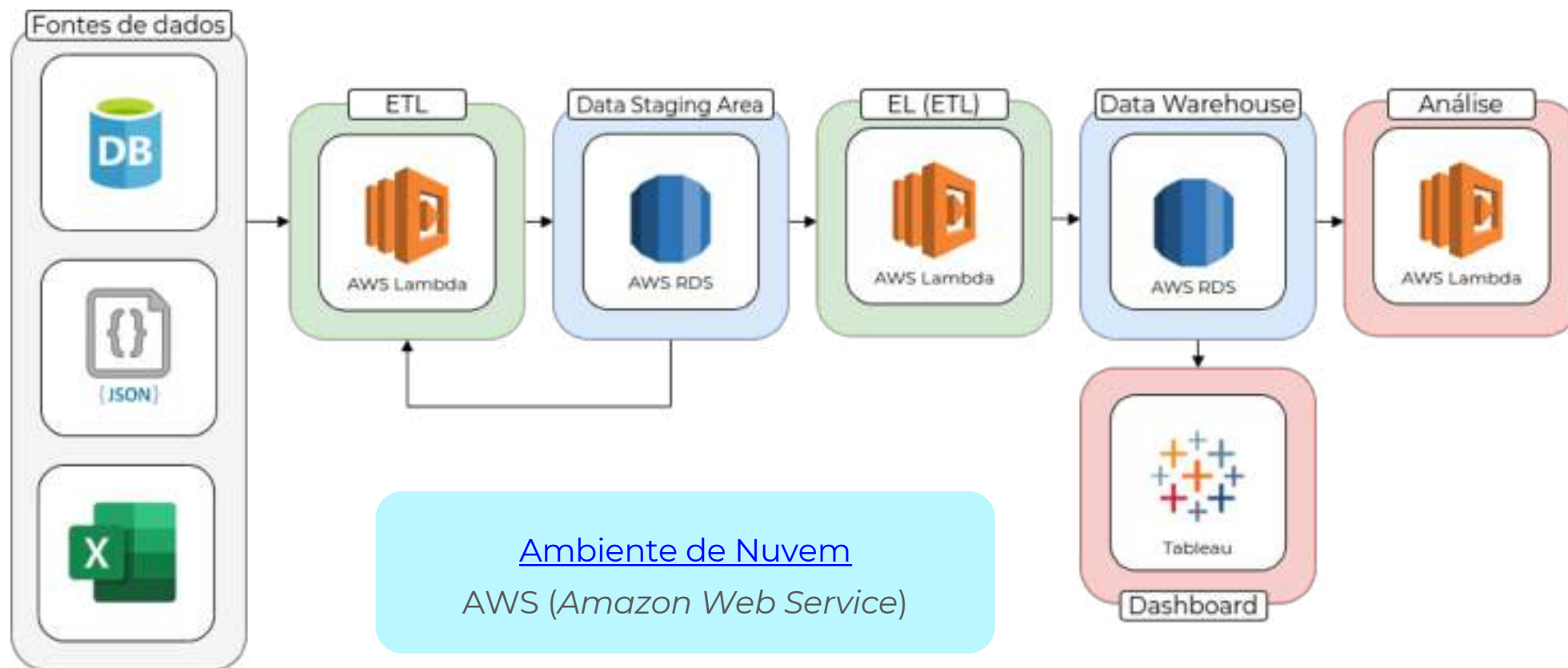
# Processamento de Dados em Lote



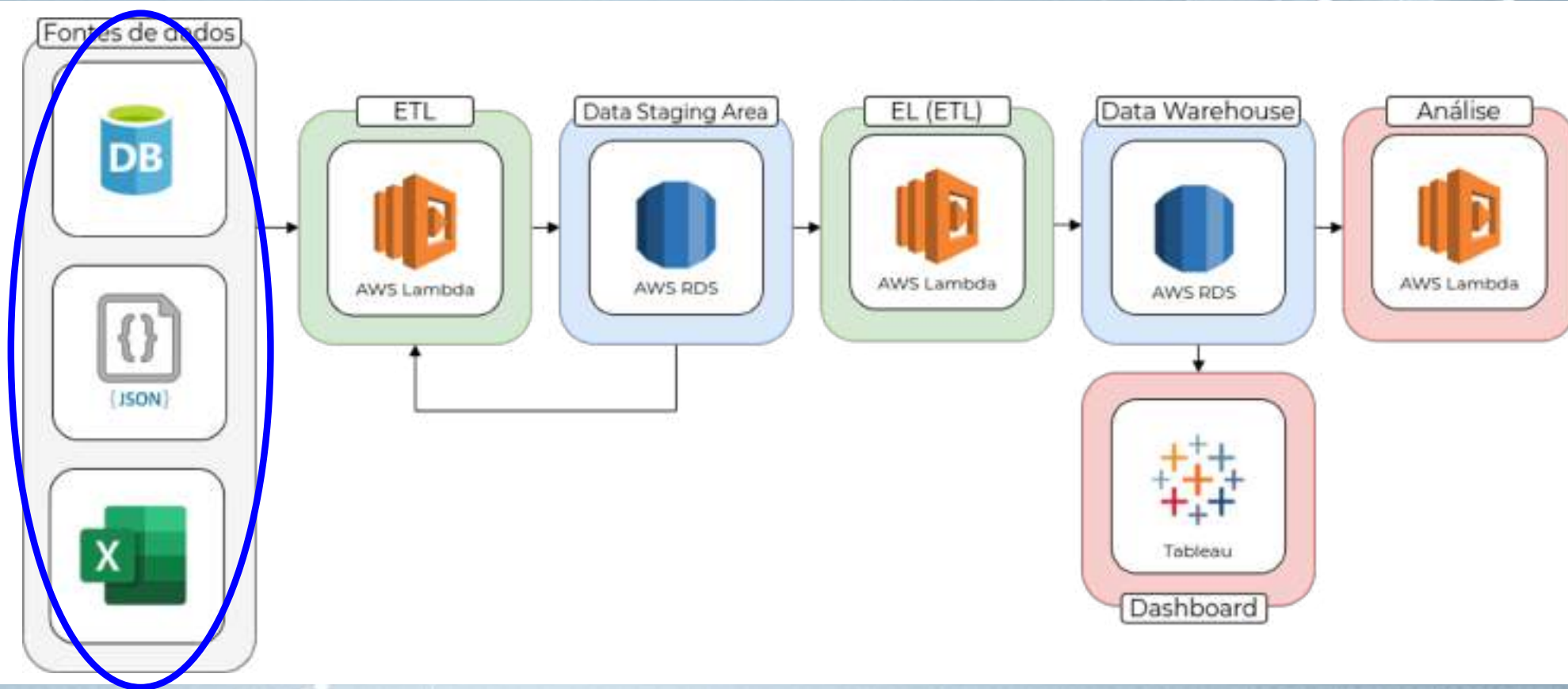
# Processamento de Dados em Lote



# Processamento de Dados em Lote (Nuvem)

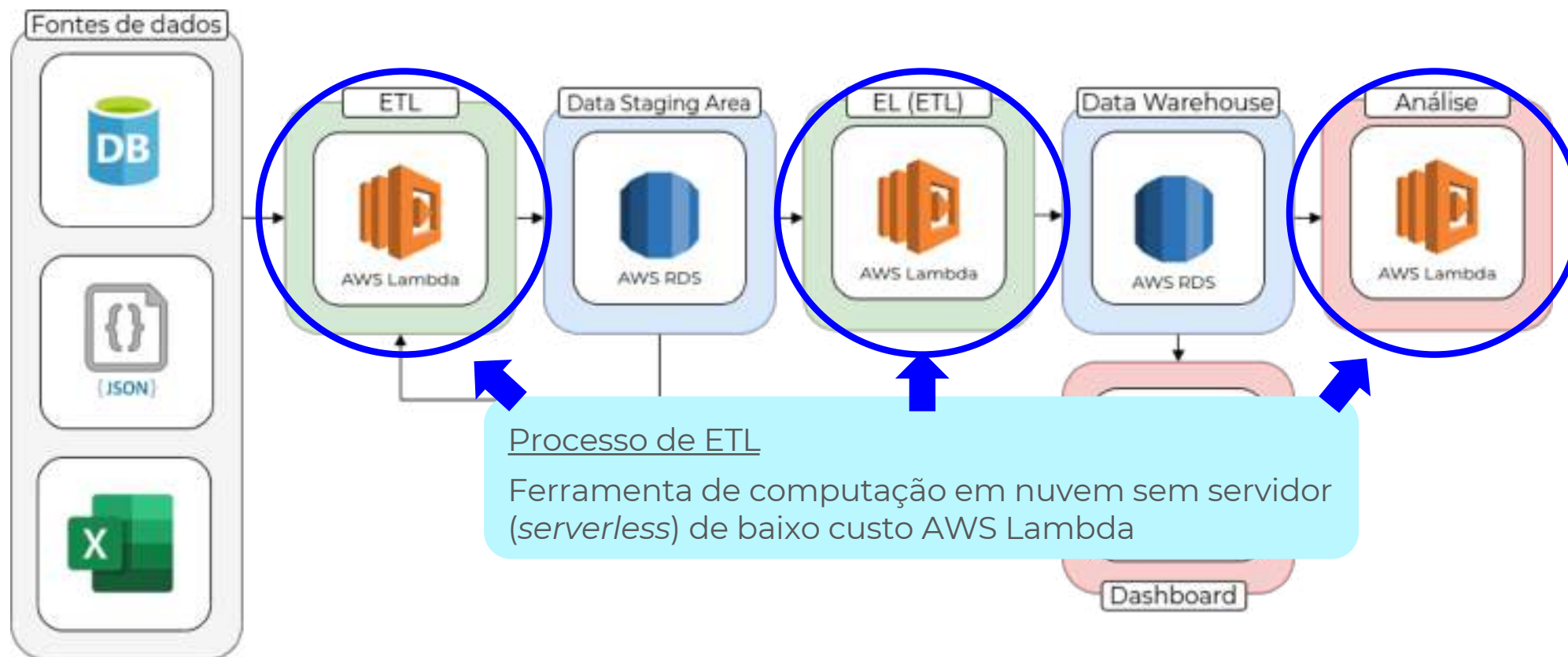


# Processamento de Dados em Lote (Nuvem)

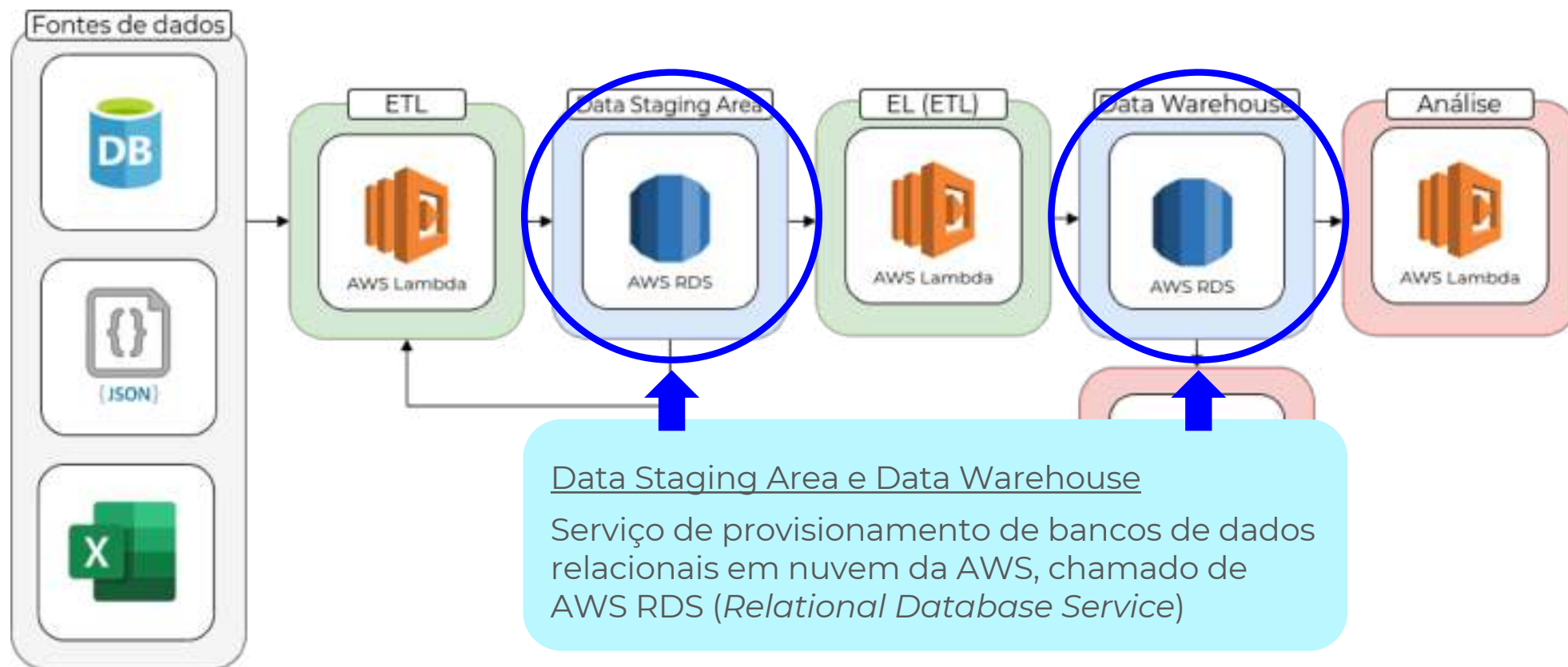




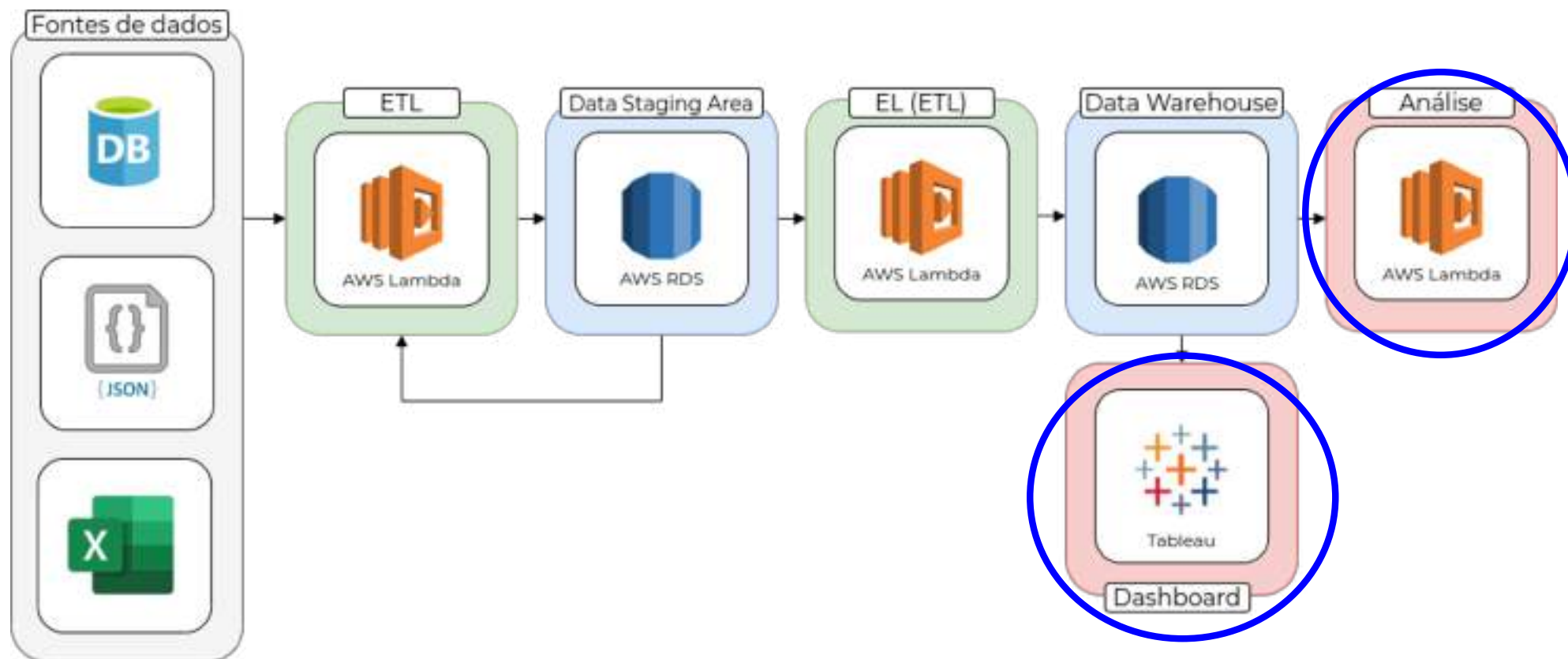
# Processamento de Dados em Lote (Nuvem)



# Processamento de Dados em Lote (Nuvem)



# Processamento de Dados em Lote (Nuvem)

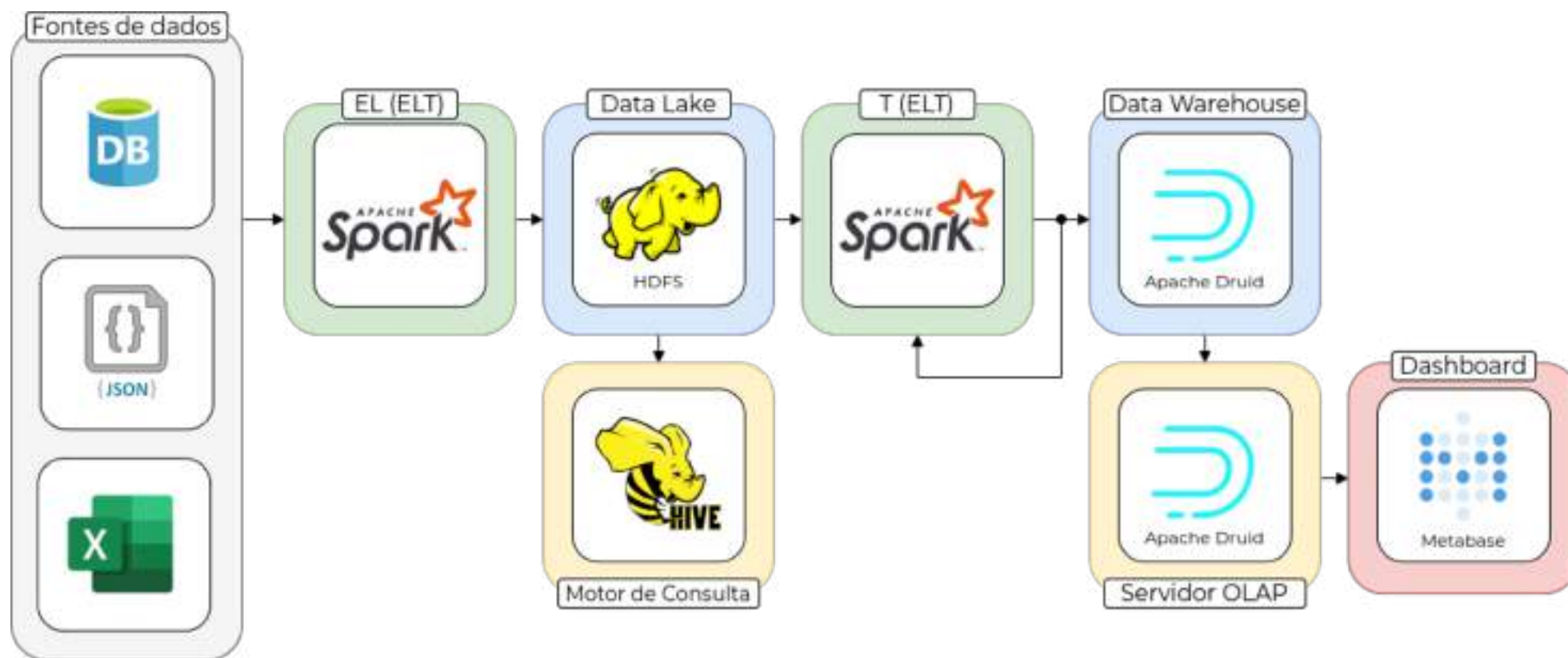




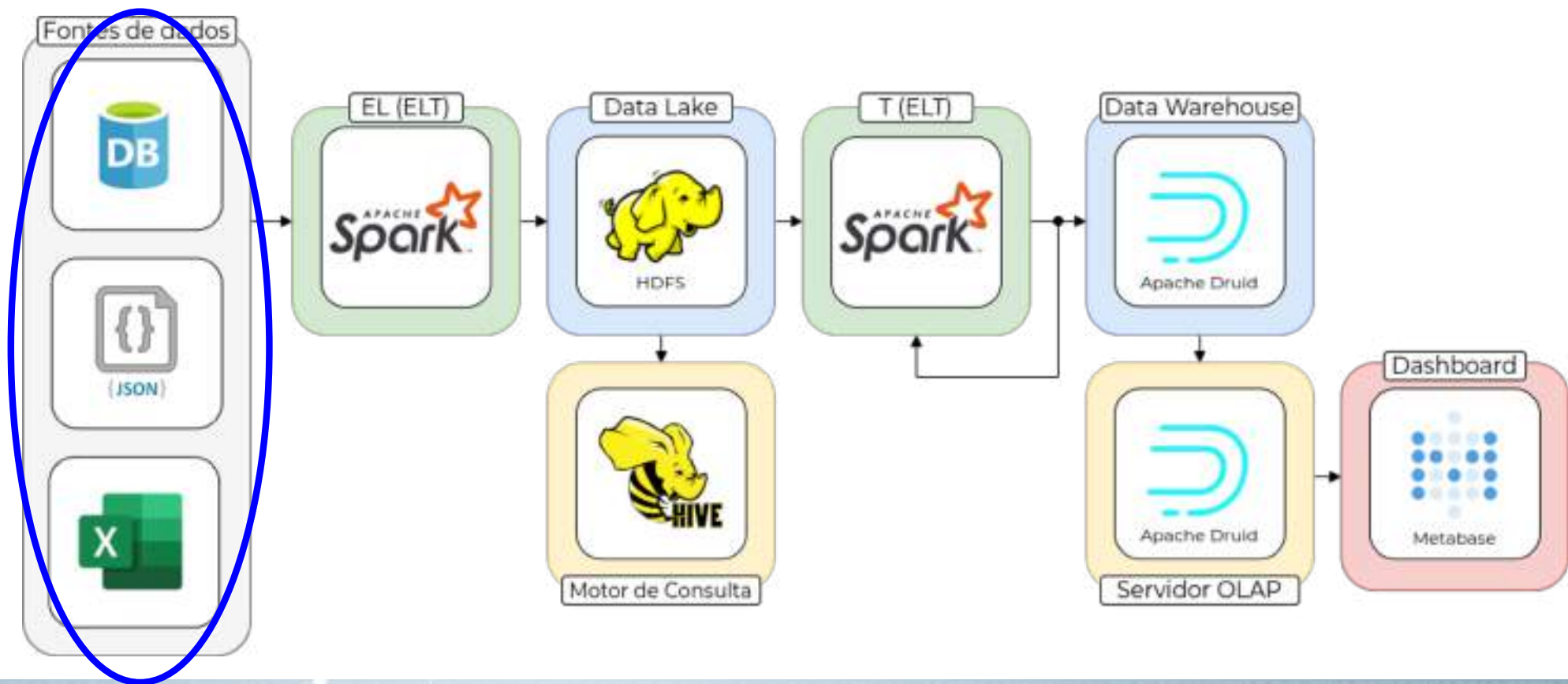
# Exemplos de Pipeline

- Volumes de Dados Tradicionais
- Big Data
- Data Streaming

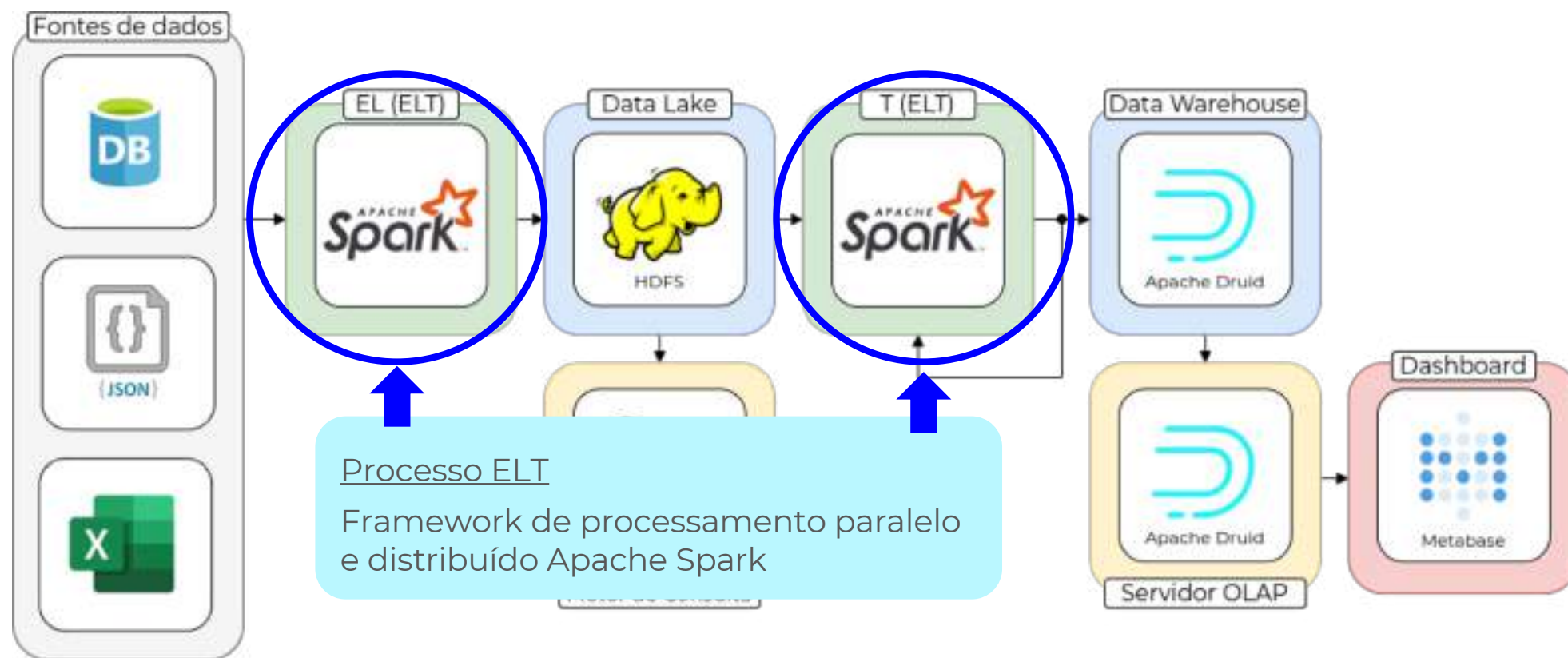
# Processamento de Big Data em Lote



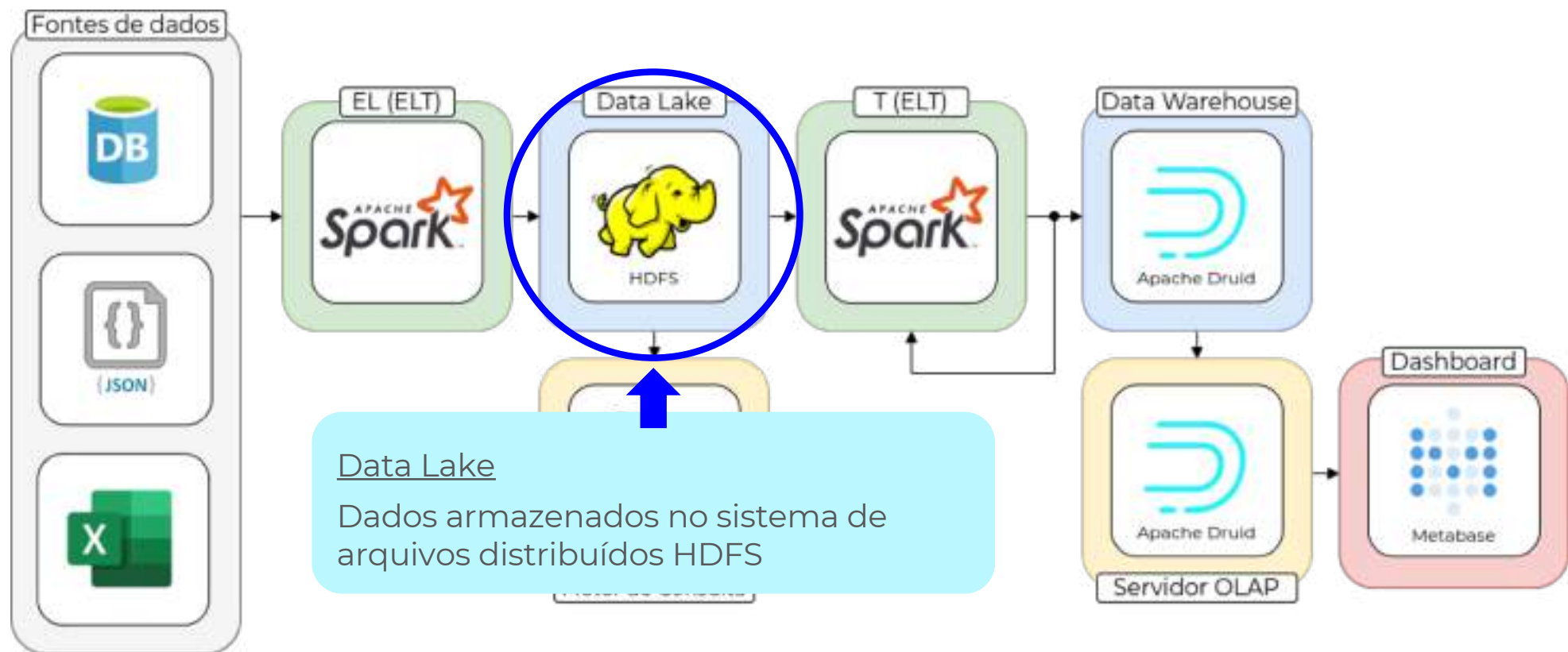
# Processamento de Big Data em Lote



# Processamento de Big Data em Lote

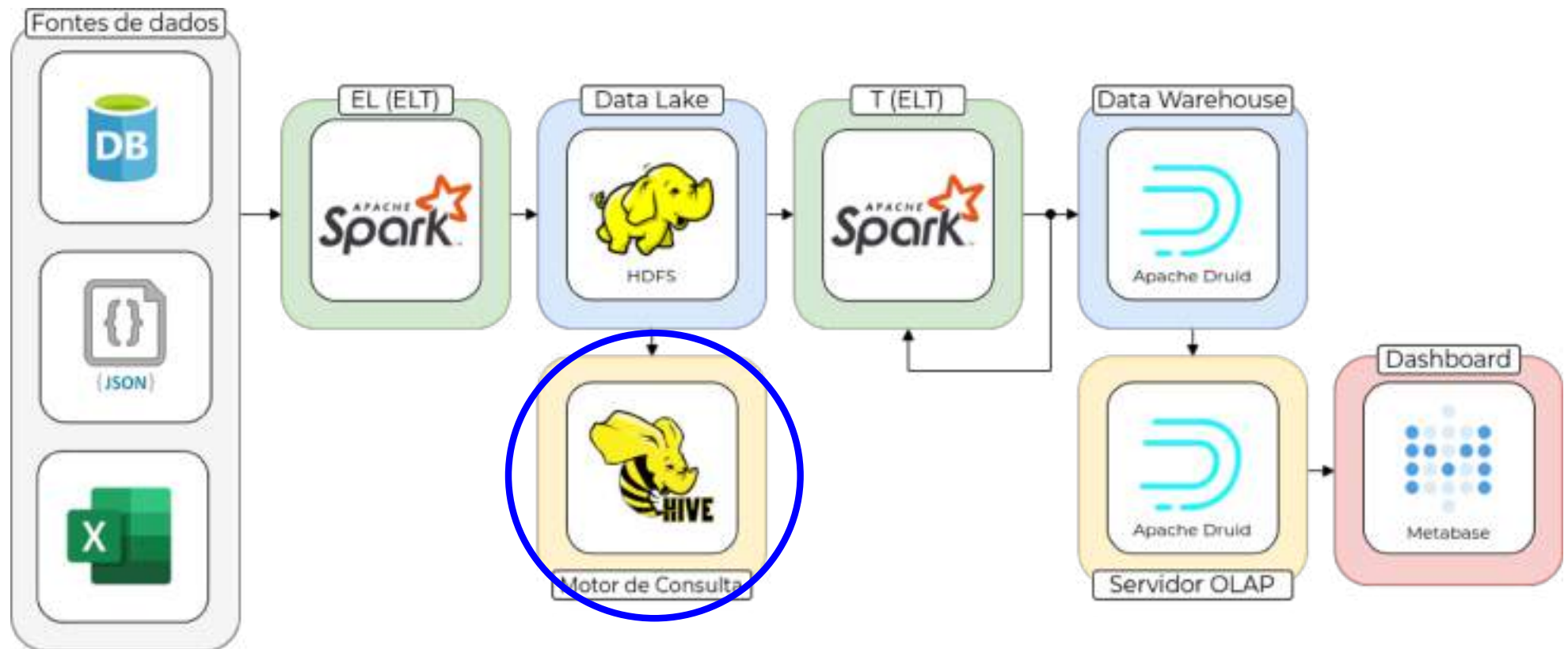


# Processamento de Big Data em Lote

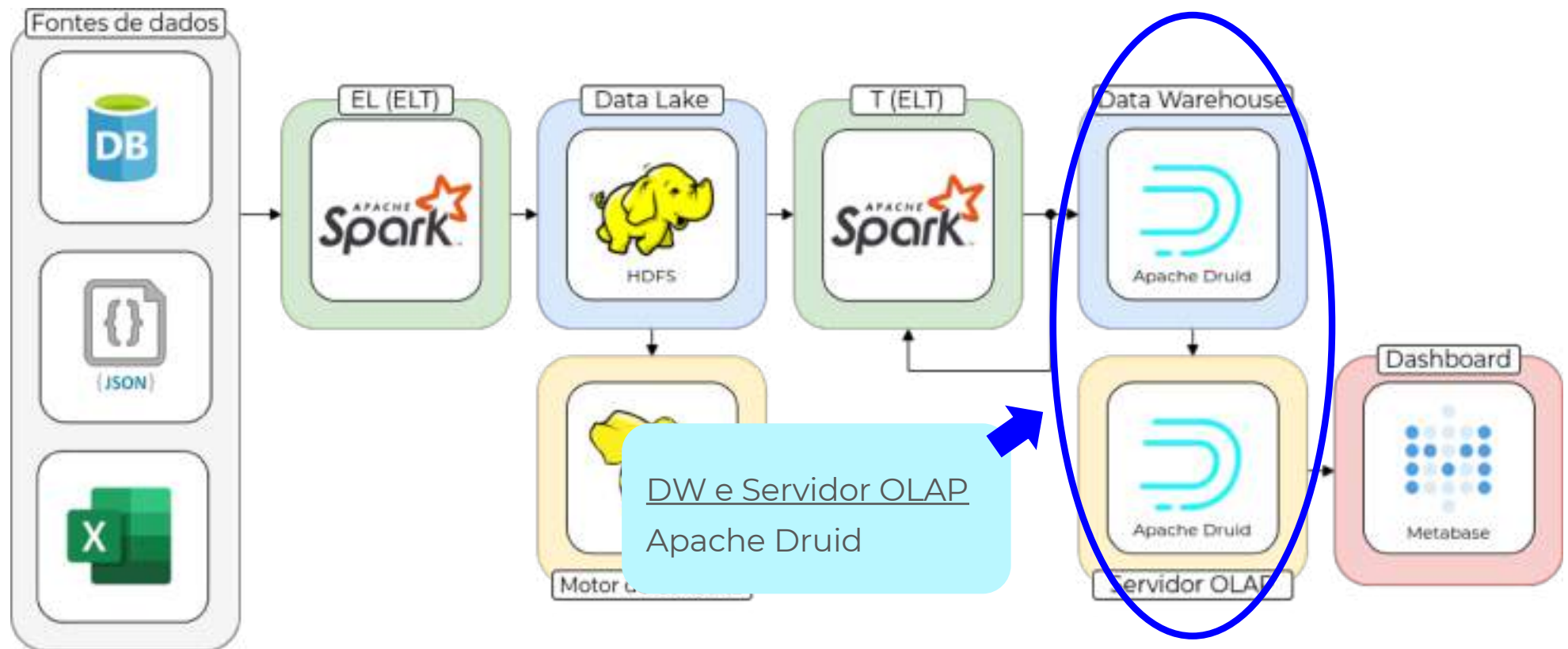




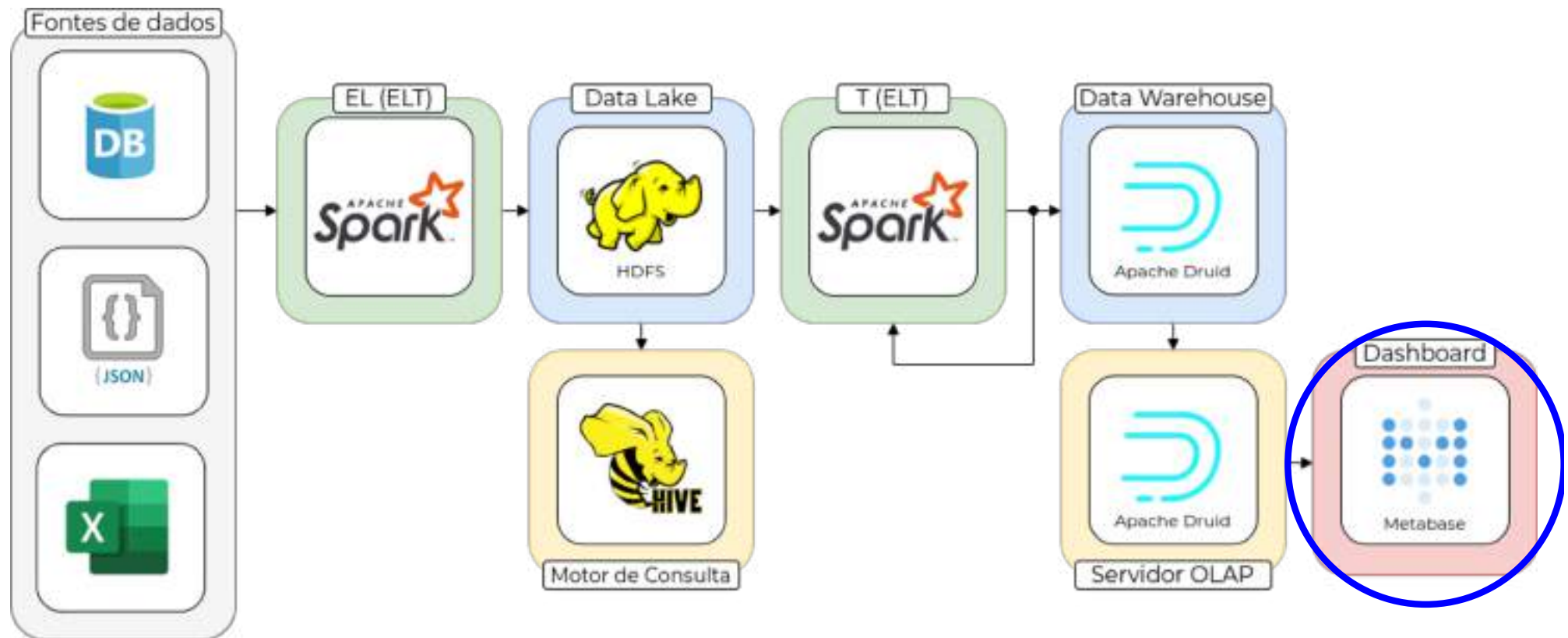
# Processamento de Big Data em Lote



# Processamento de Big Data em Lote

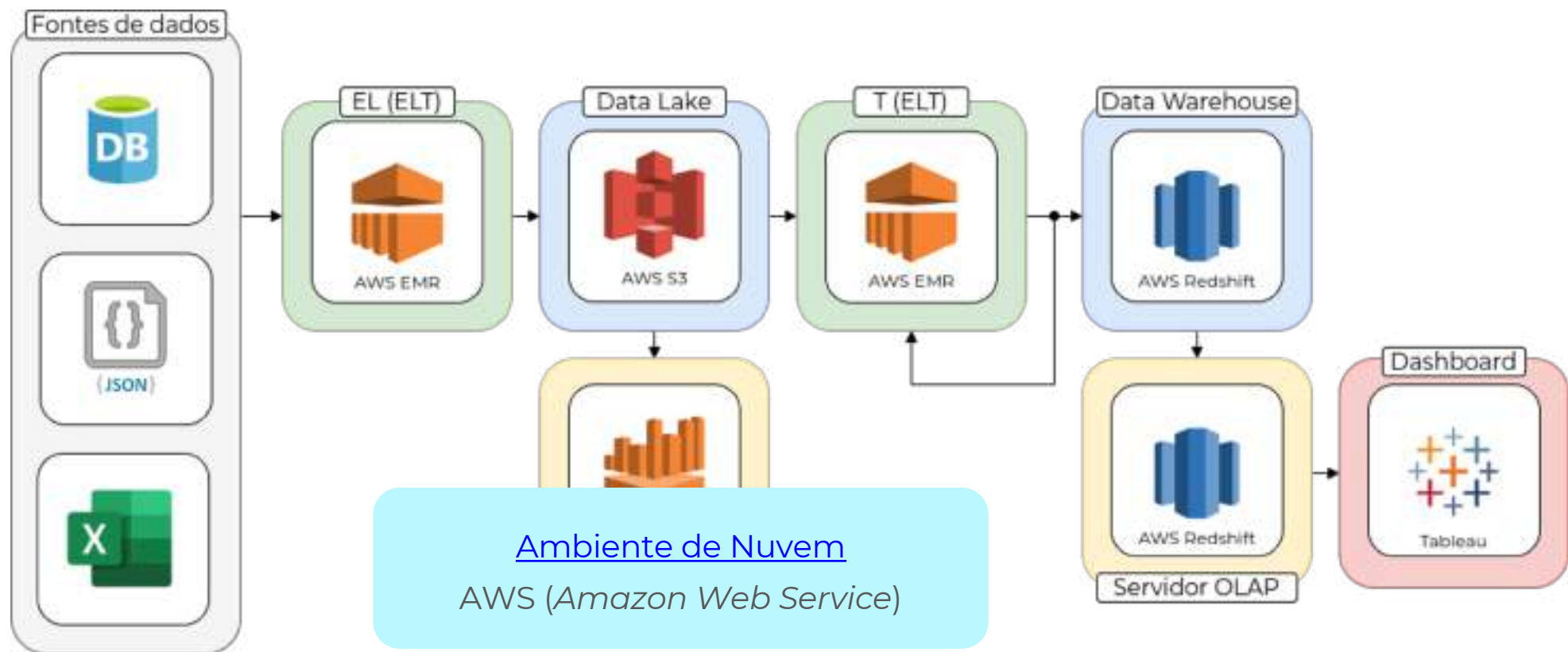


# Processamento de Big Data em Lote

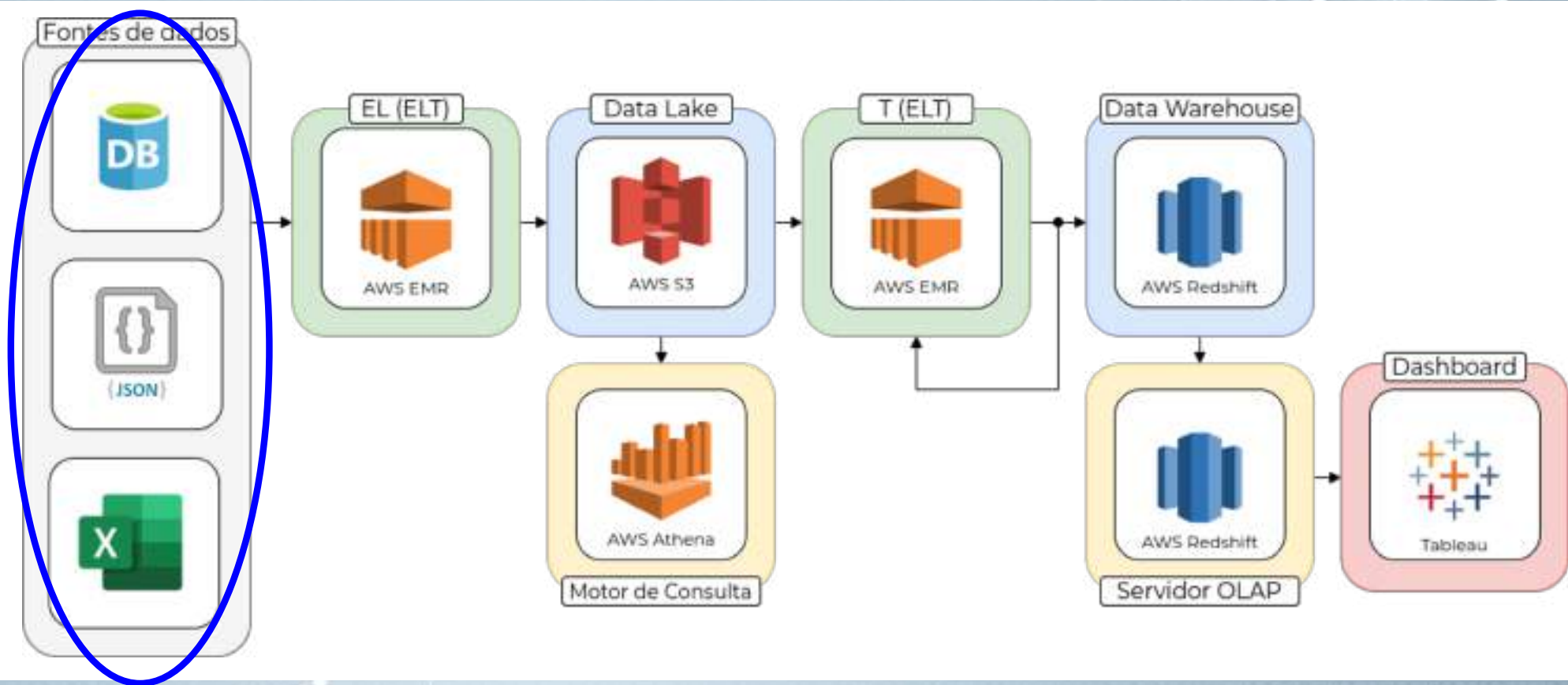




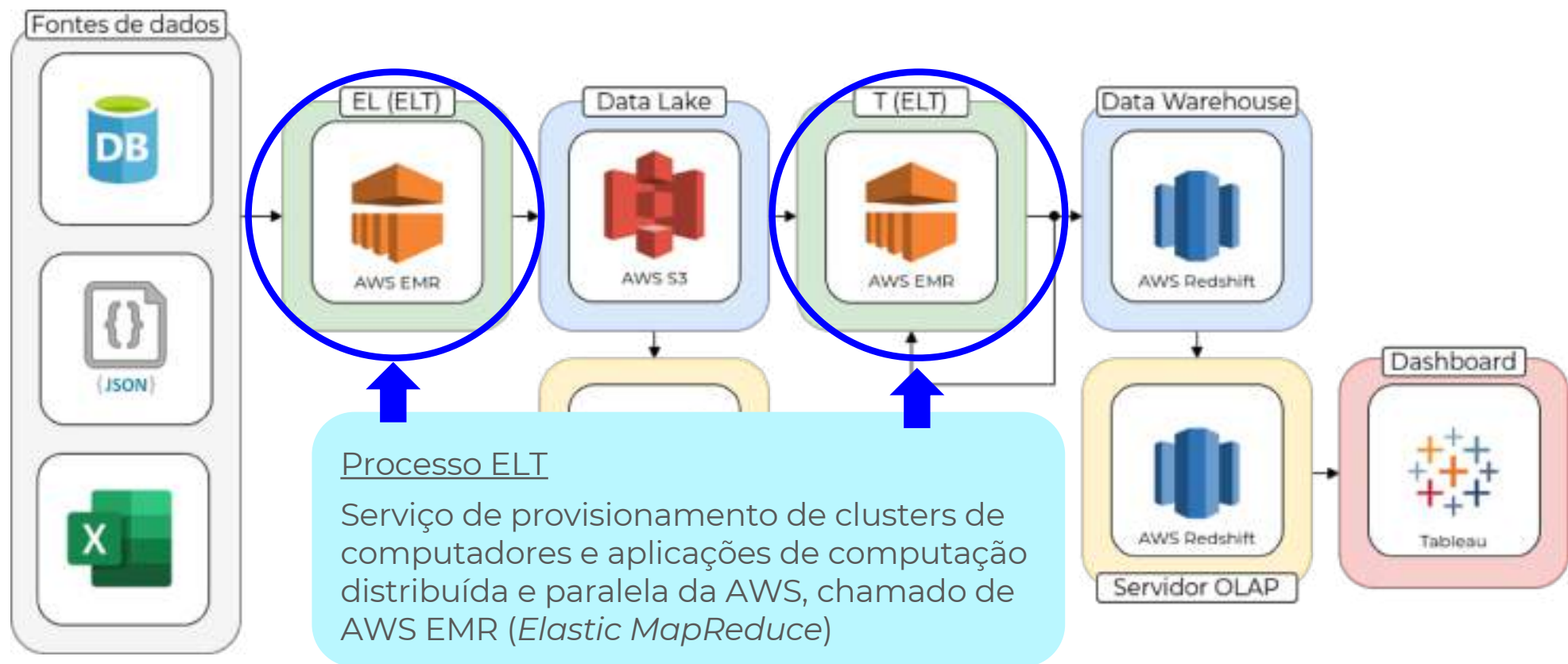
# Processamento de Big Data em Lote (Nuvem)



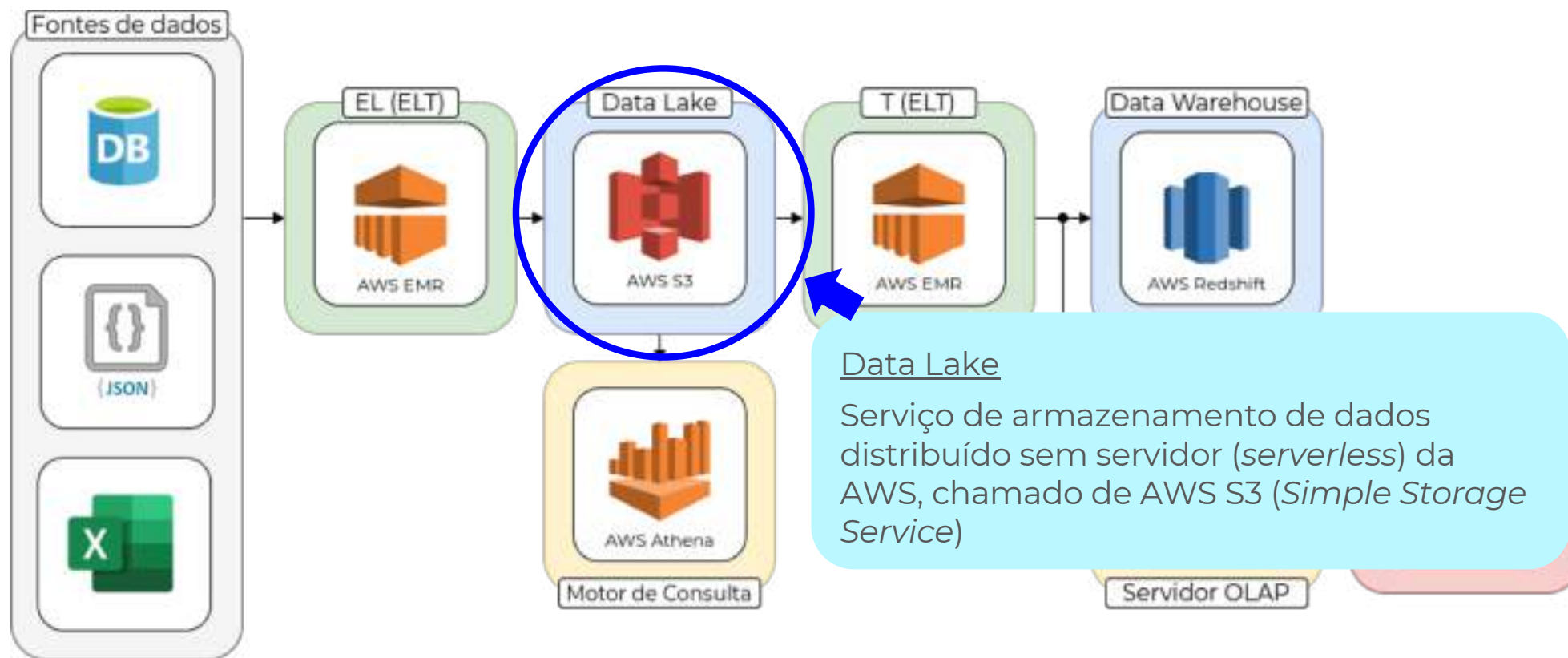
# Processamento de Big Data em Lote (Nuvem)



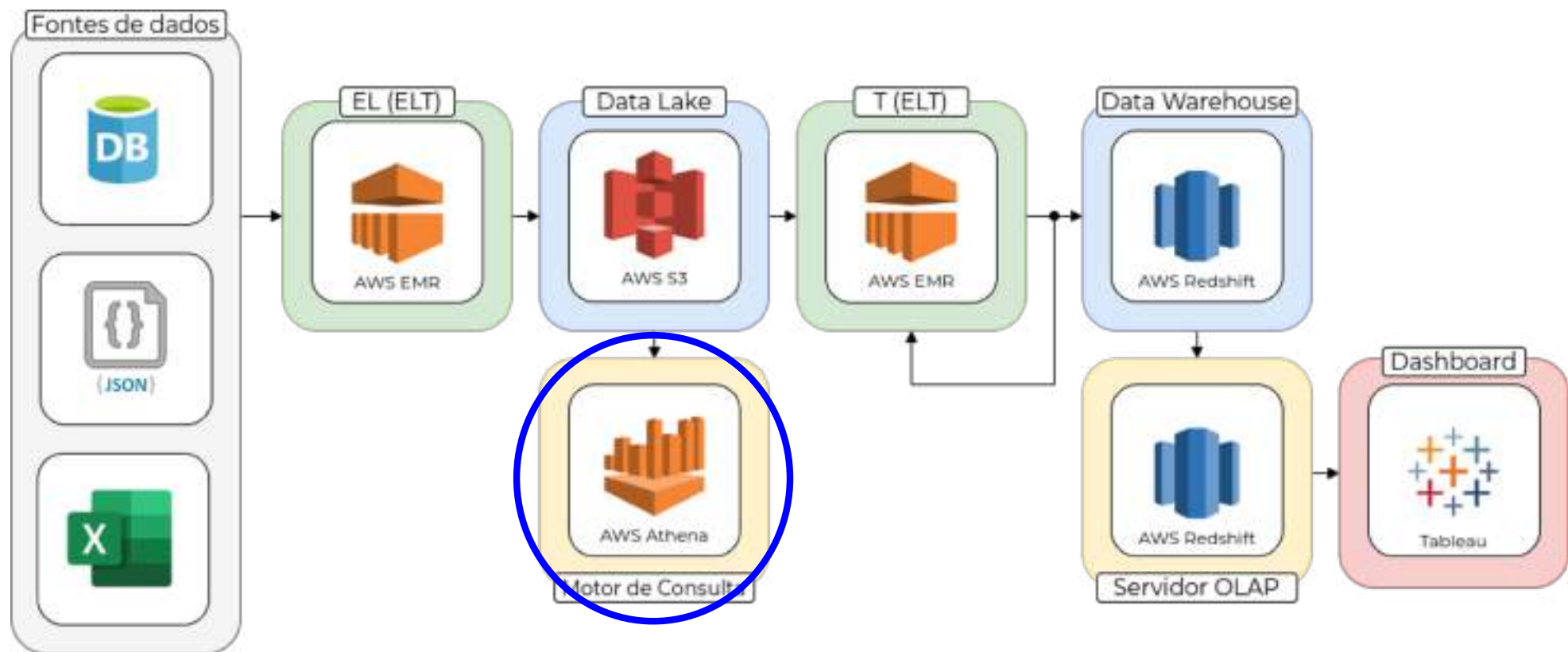
# Processamento de Big Data em Lote (Nuvem)



# Processamento de Big Data em Lote (Nuvem)

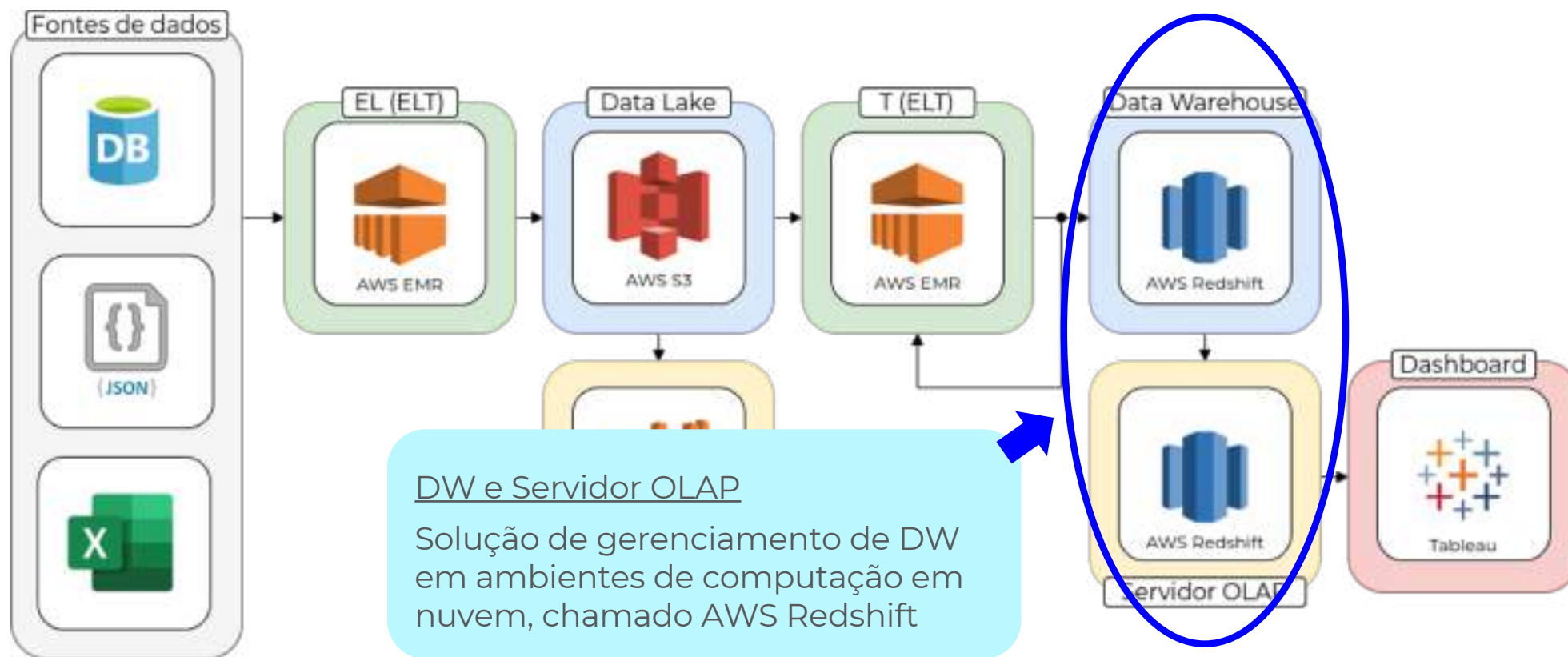


# Processamento de Big Data em Lote (Nuvem)

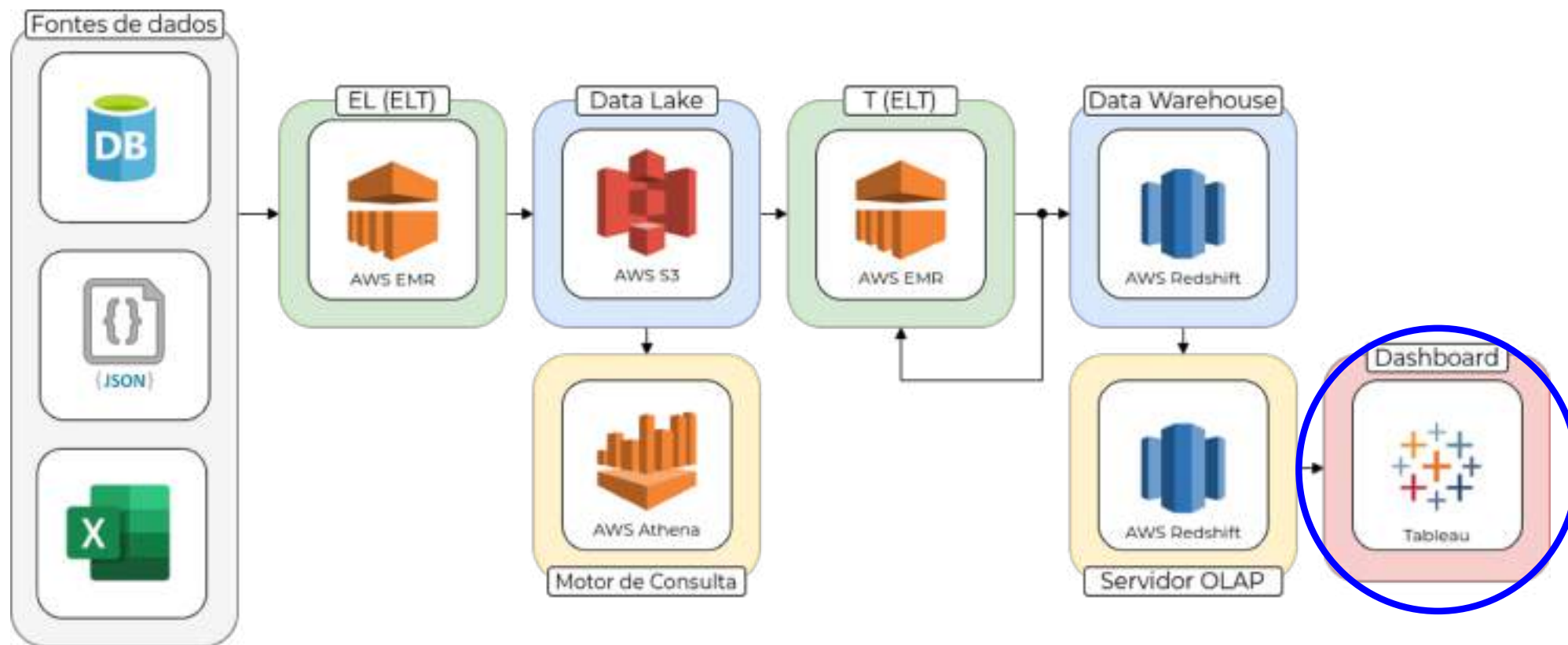




# Processamento de Big Data em Lote (Nuvem)



# Processamento de Big Data em Lote (Nuvem)

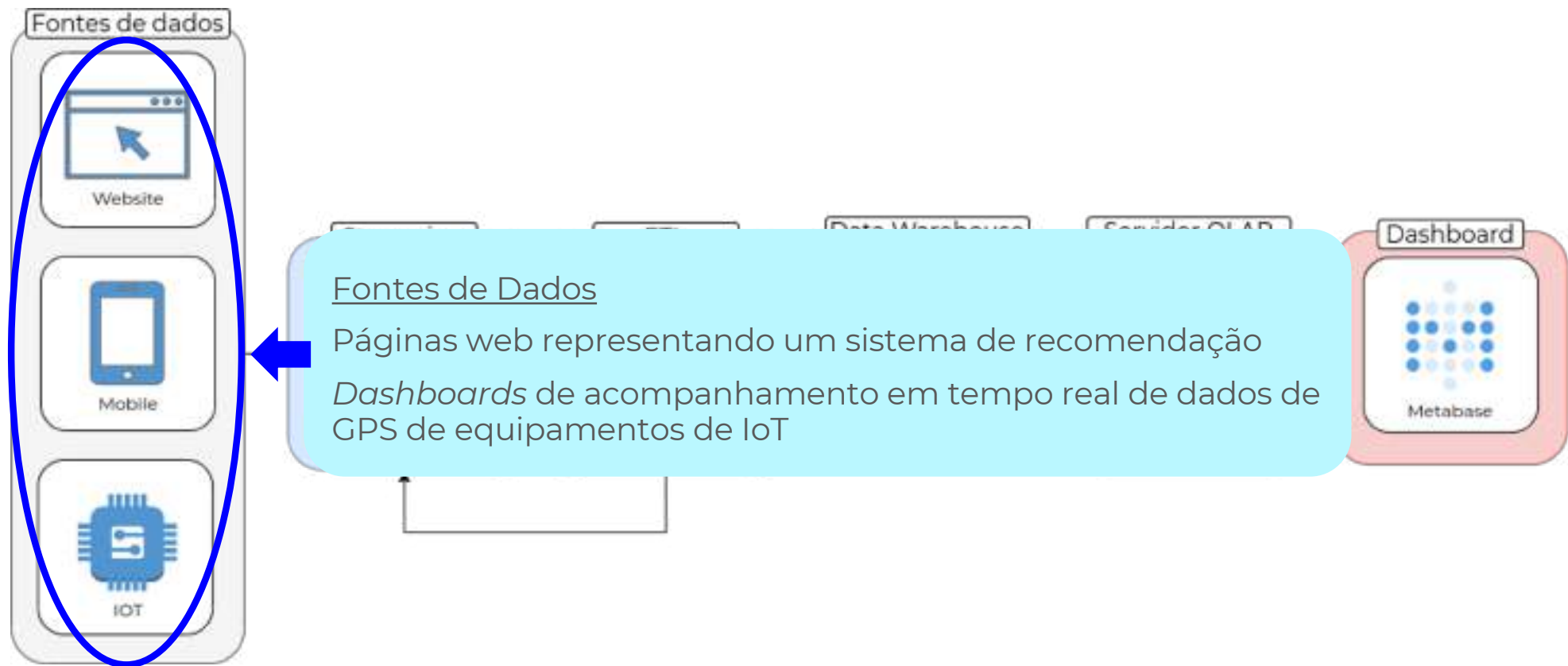


# Exemplos de Pipeline

- Volumes de Dados Tradicionais
- Big Data
- Data Streaming



# Processamento de Streaming de Big Data



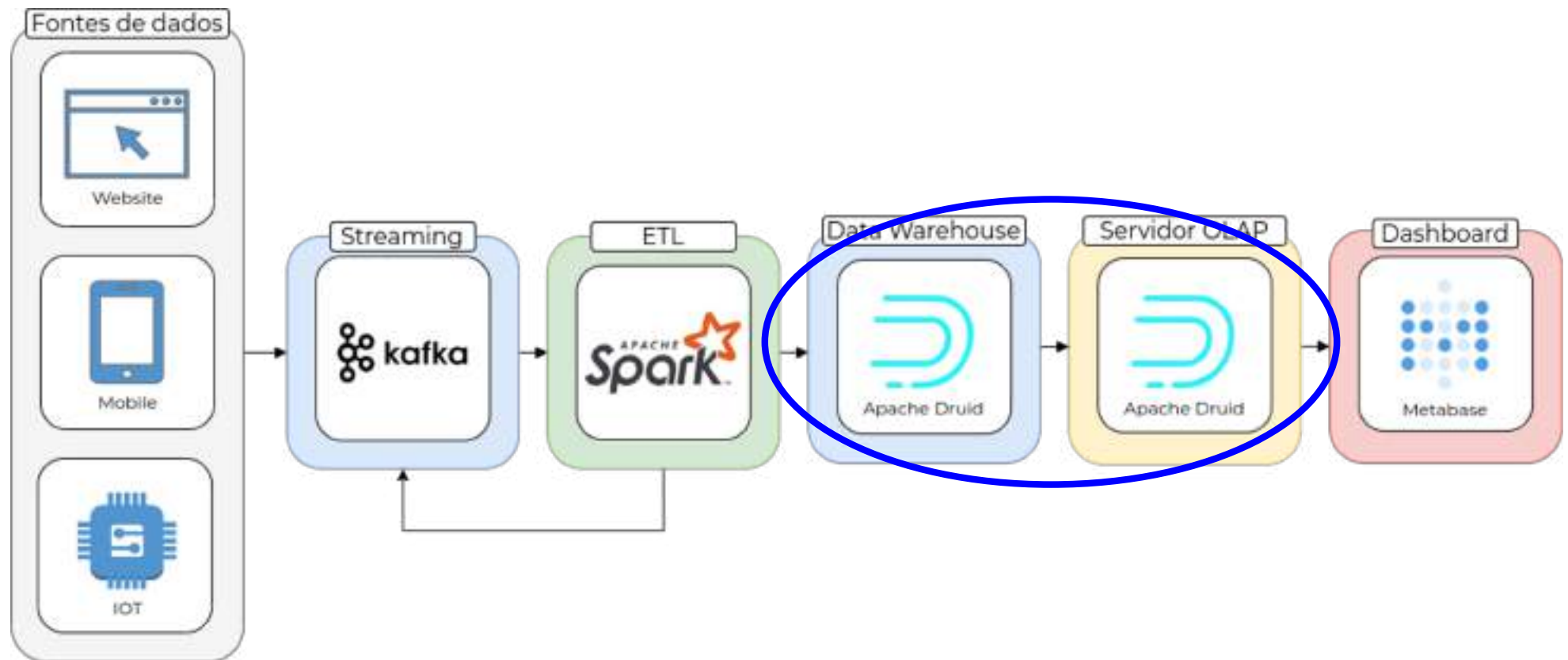
# Processamento de Streaming de Big Data



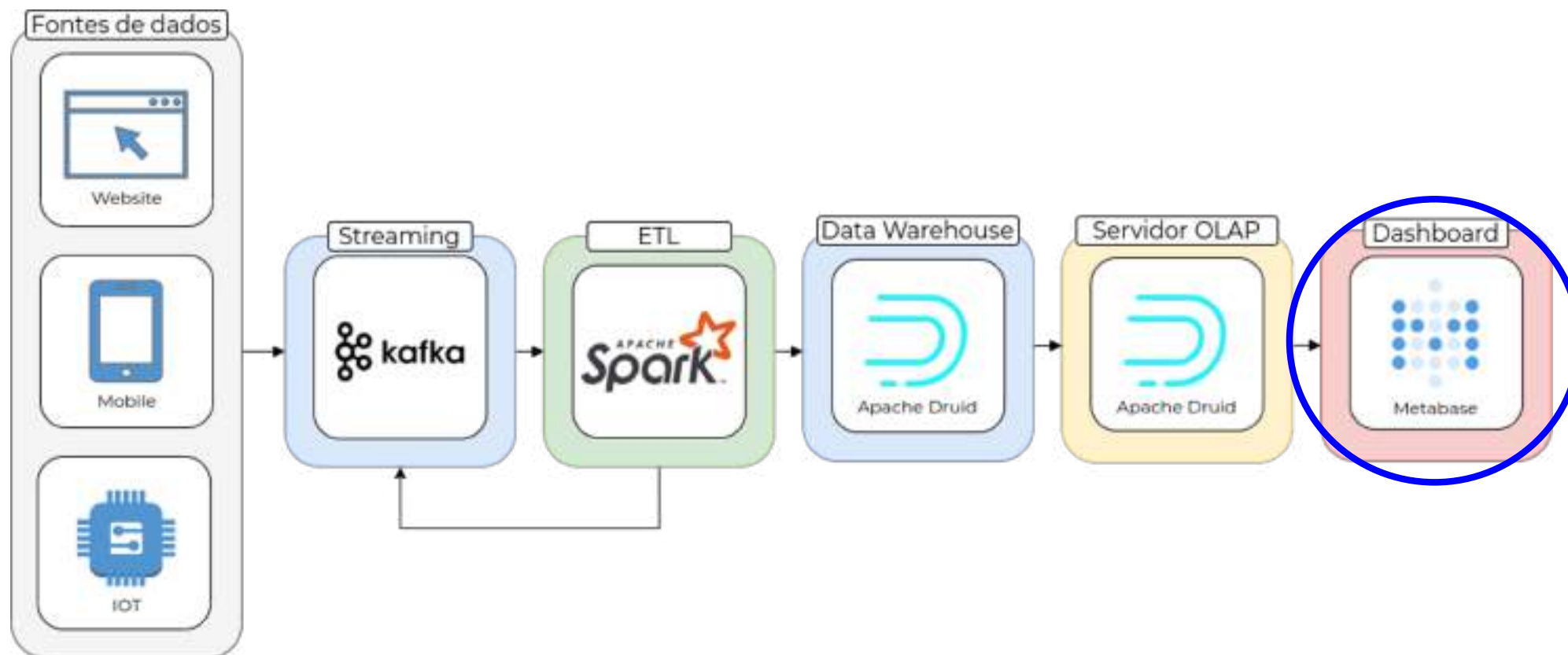
# Processamento de Streaming de Big Data



# Processamento de Streaming de Big Data

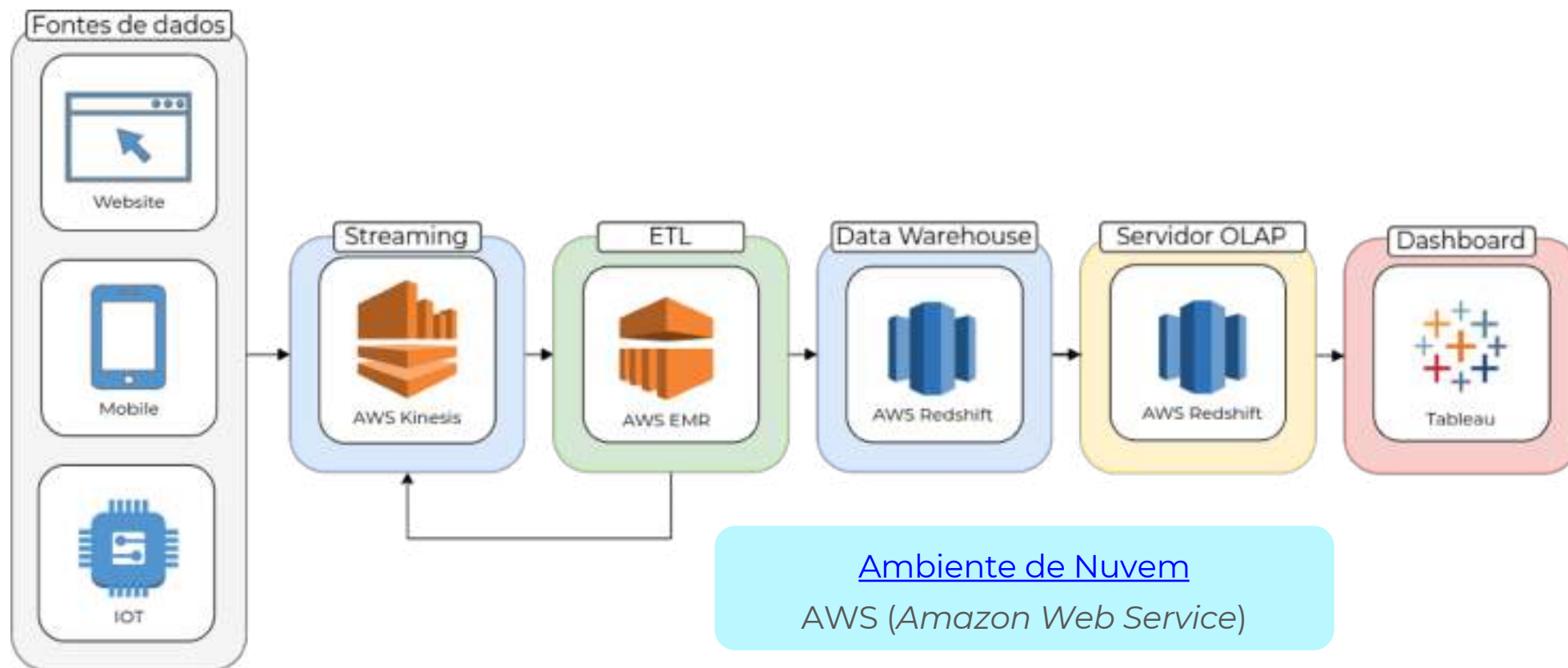


# Processamento de Streaming de Big Data

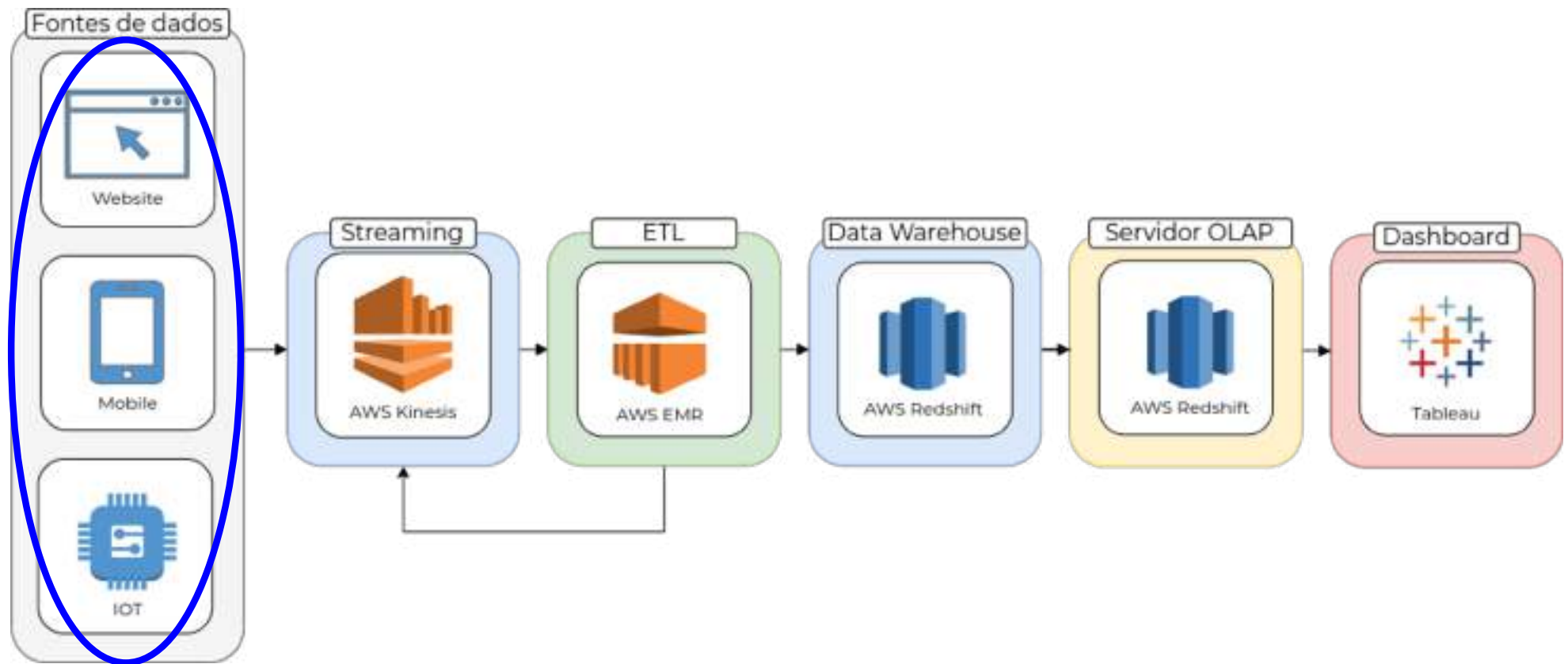




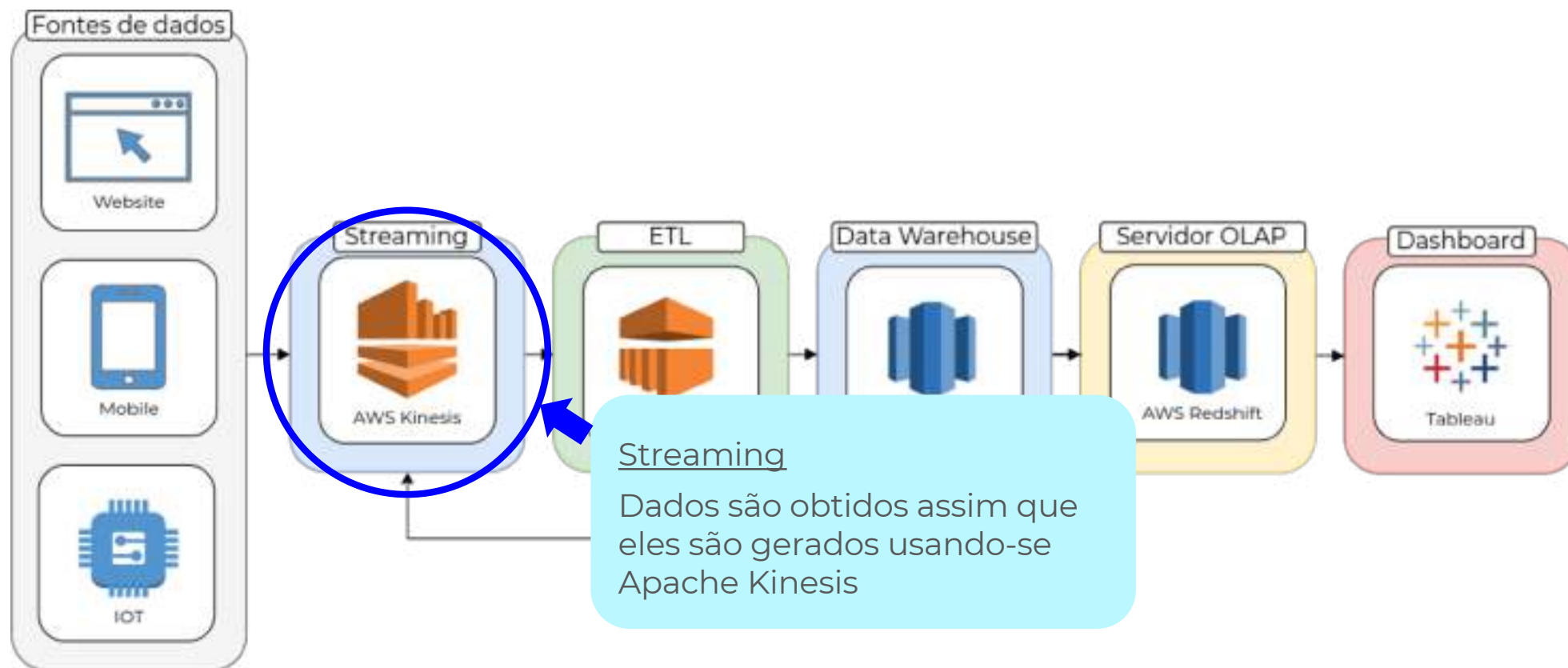
# Processamento de Streaming de Big Data (Nuvem)



# Processamento de Streaming de Big Data (Nuvem)

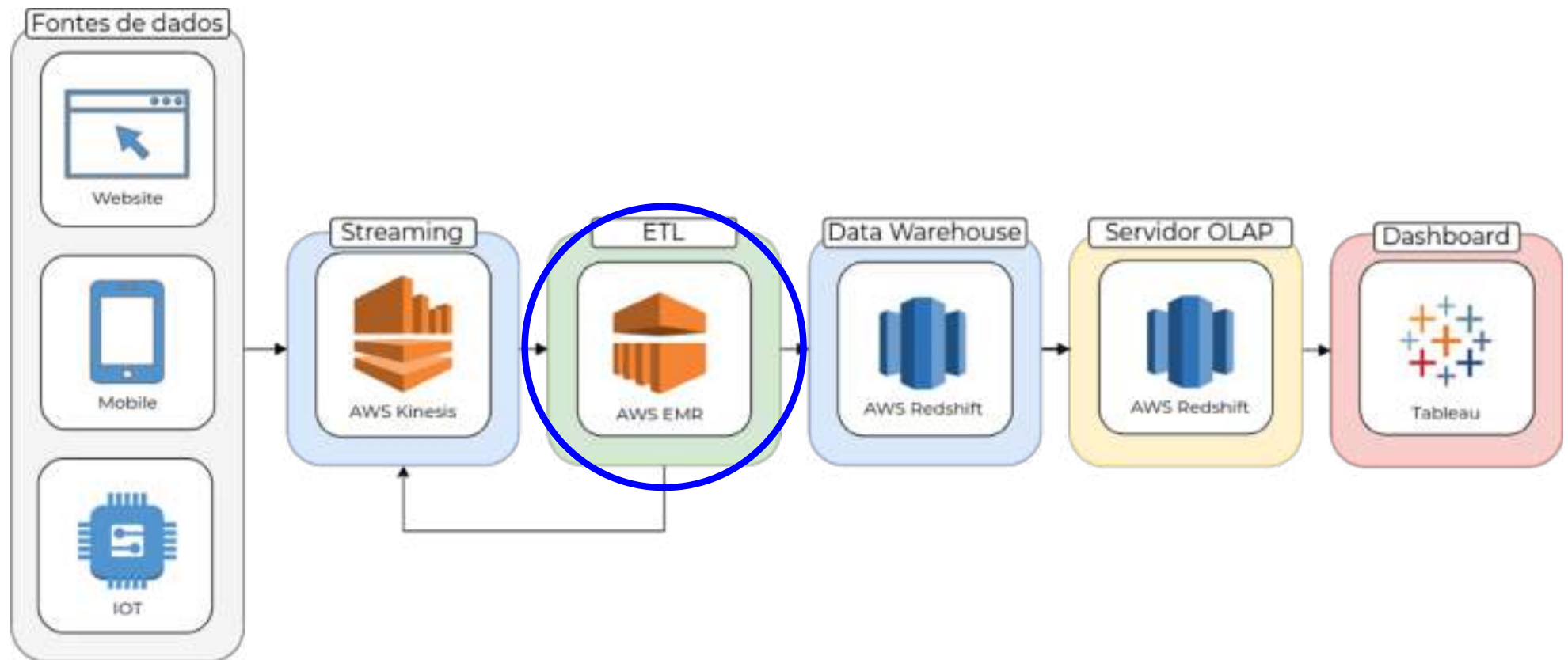


# Processamento de Streaming de Big Data (Nuvem)

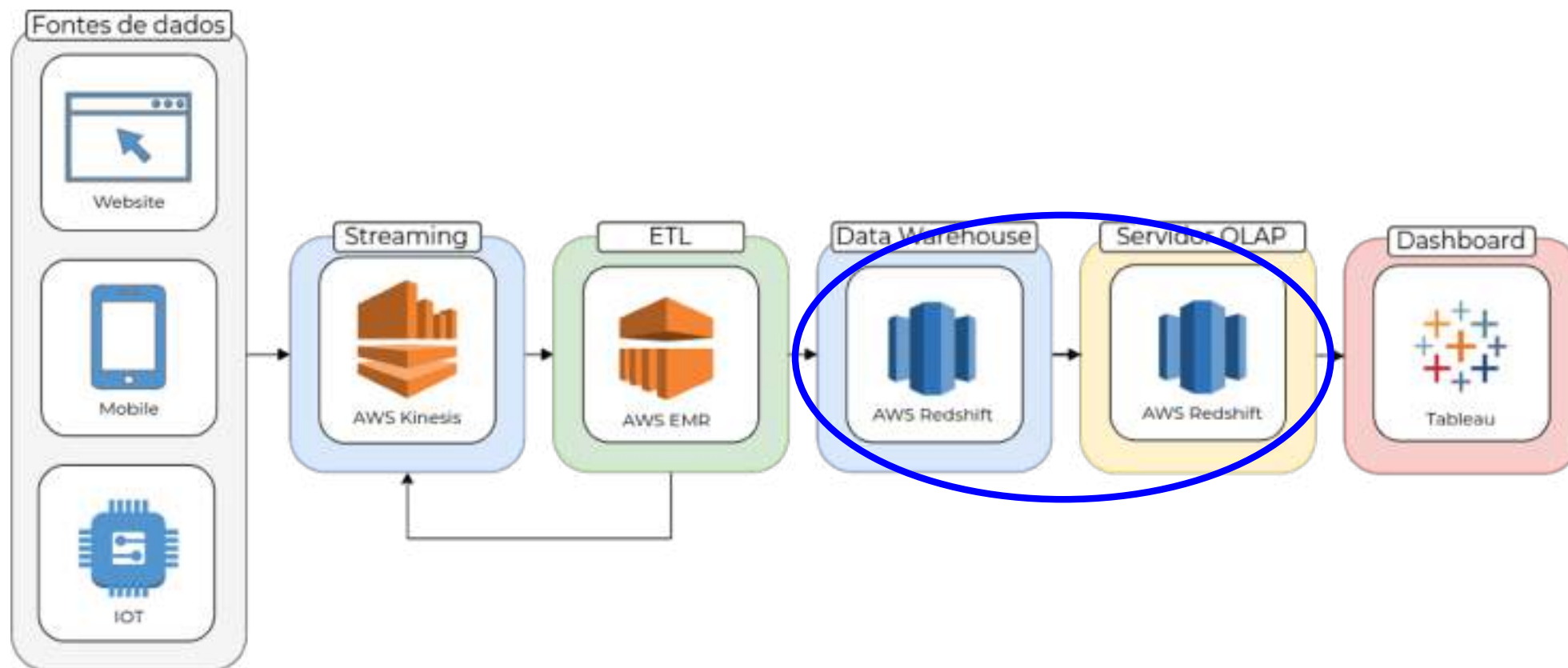




# Processamento de Streaming de Big Data (Nuvem)



# Processamento de Streaming de Big Data (Nuvem)



# Processamento de Streaming de Big Data (Nuvem)

