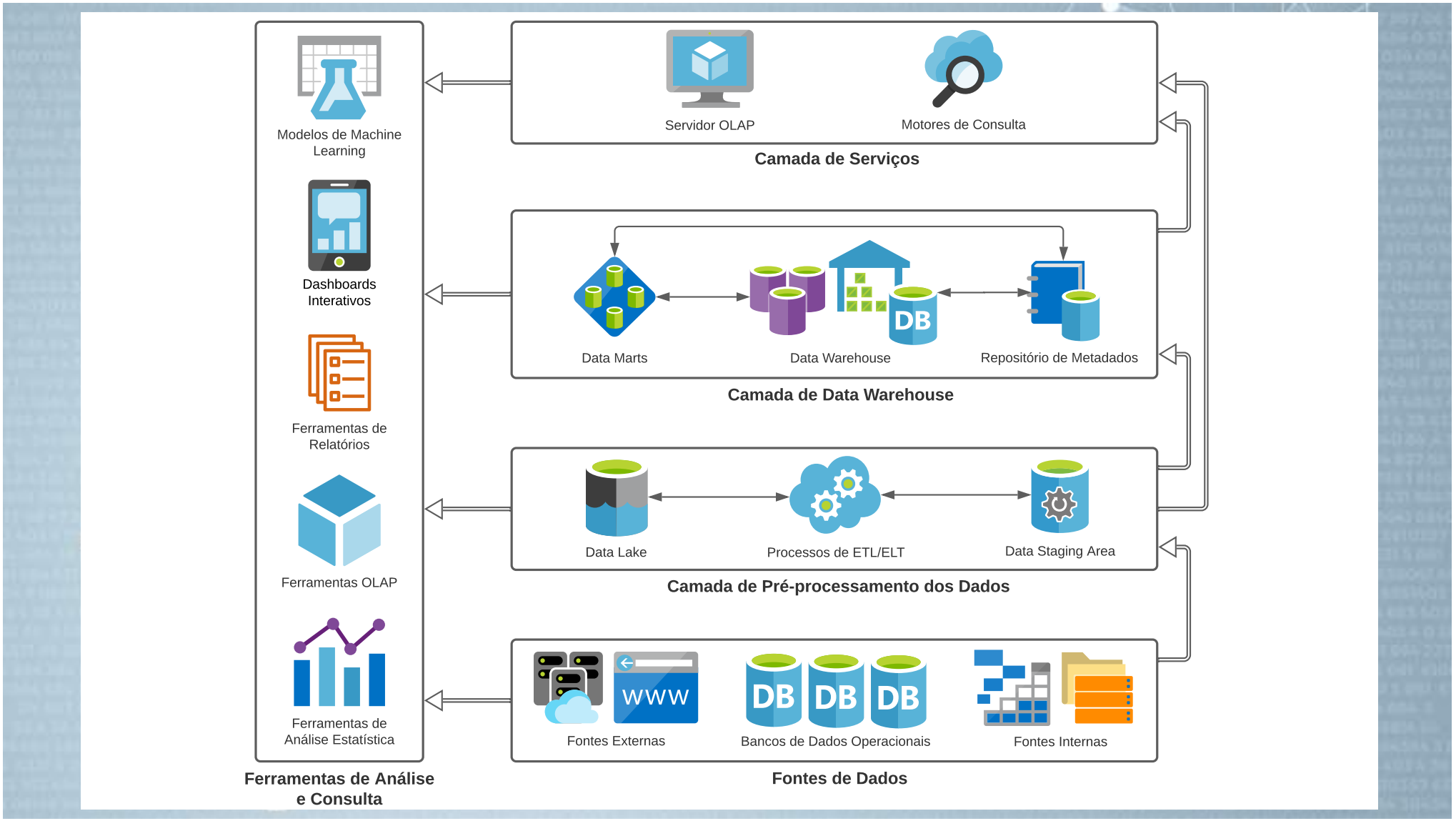
Análise de Dados com Base em Processamento Massivo em Paralelo

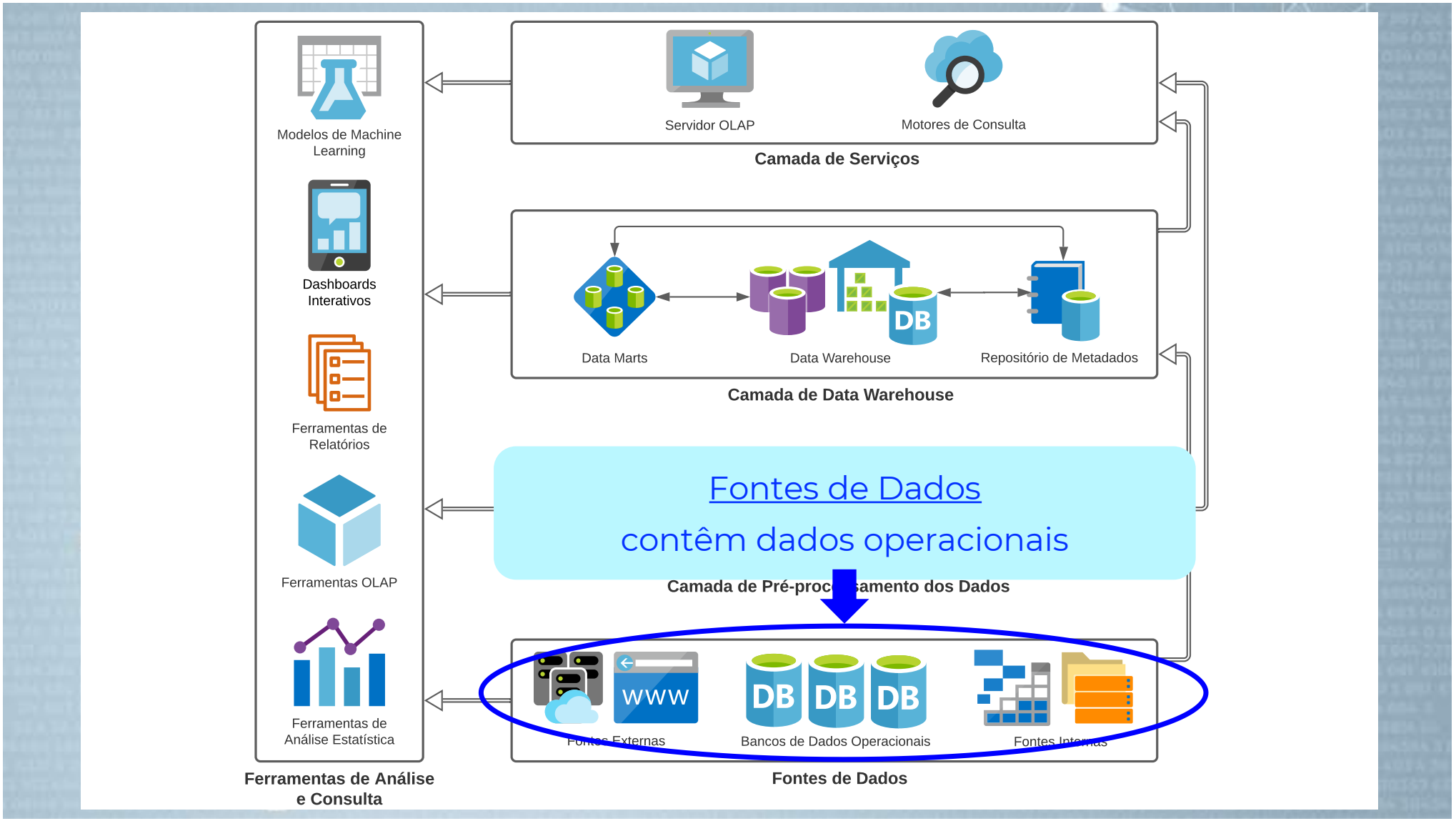
Aula 2: Arquitetura de Data Warehousing

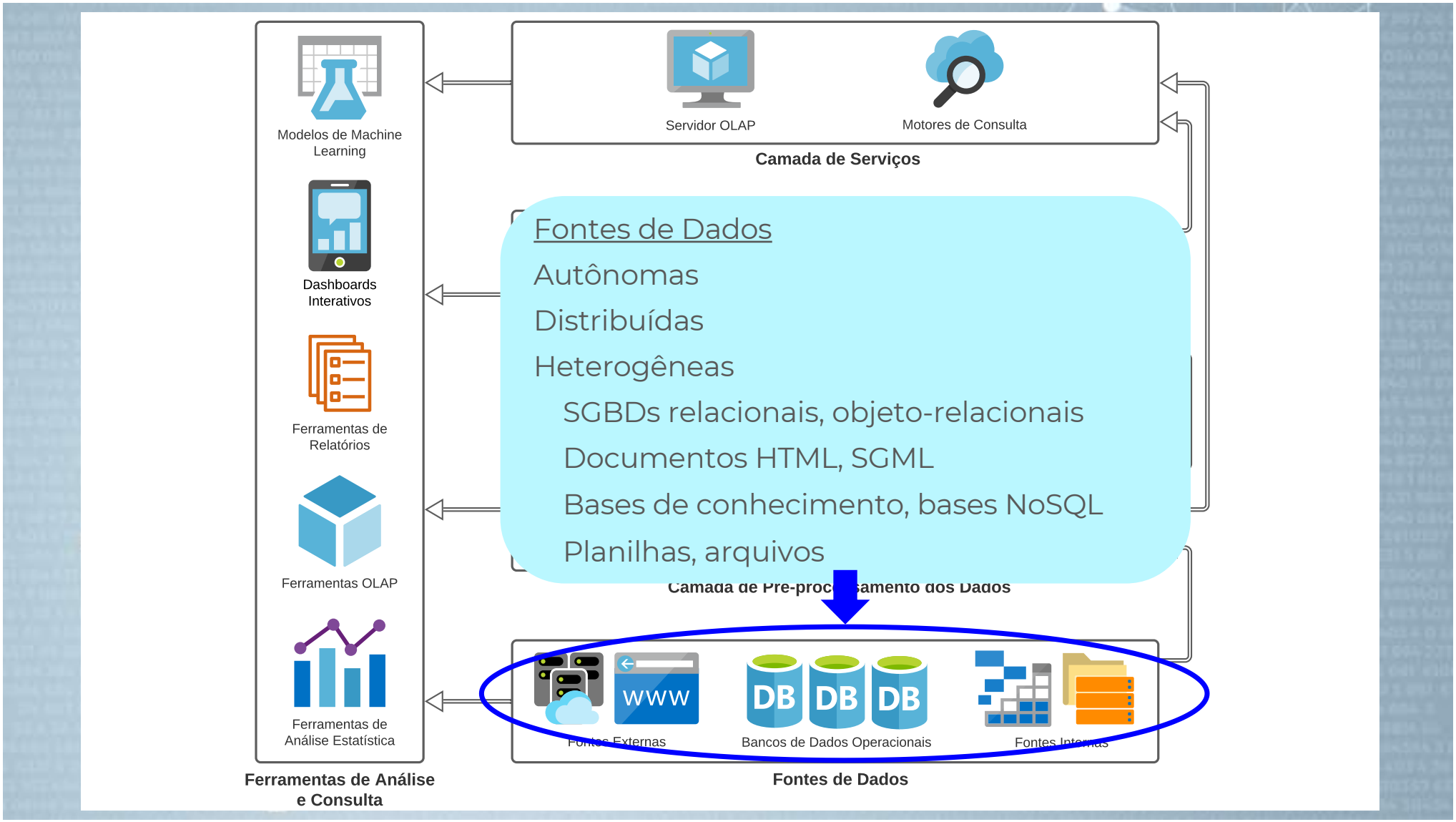
Cristina Dutra de Aguiar
ICMC/USP
cdac@icmc.usp.br

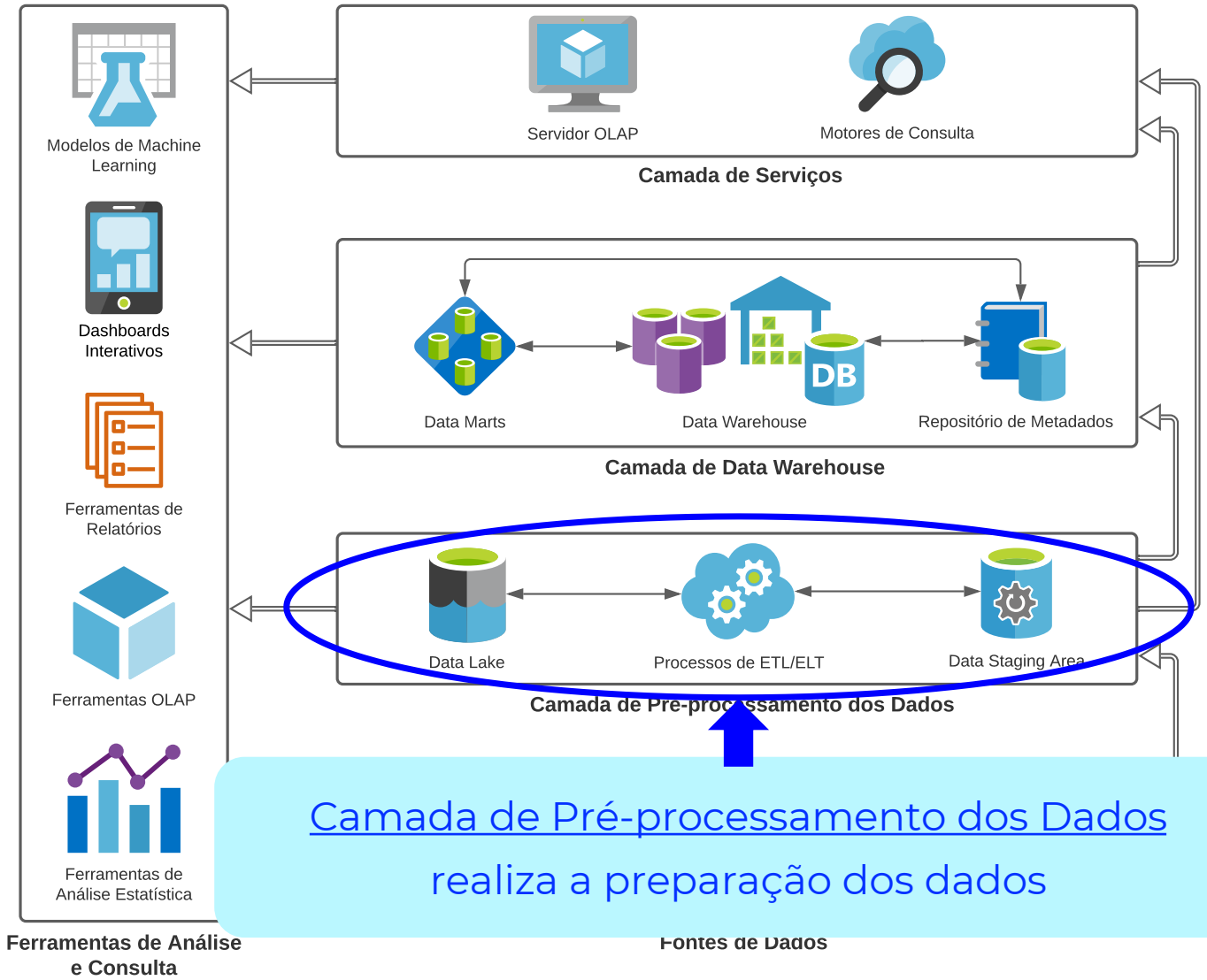
Agenda

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

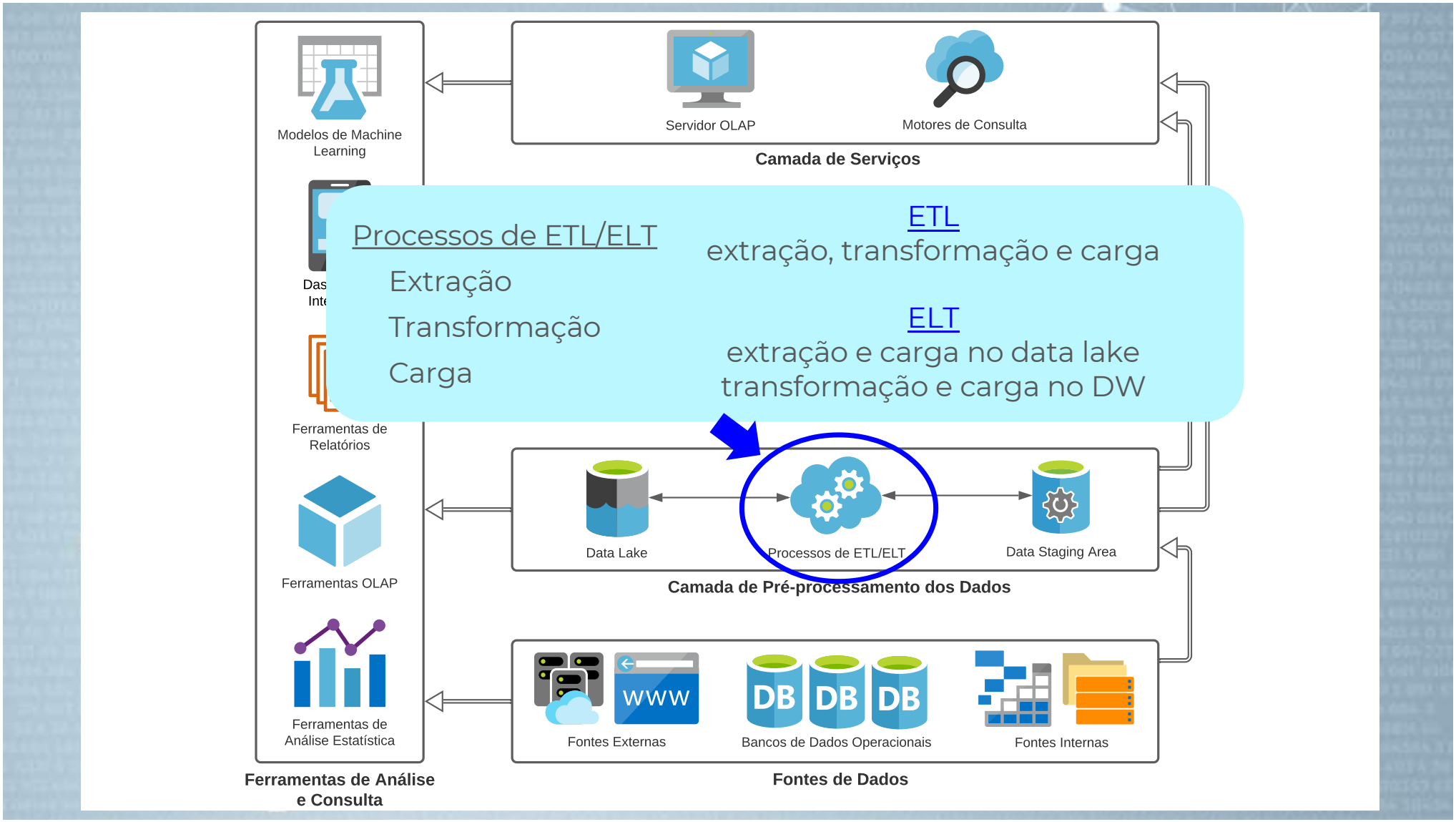

Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Data Marts
Data Warehouse
DB
Repositório de Metadados
Camada de Data Warehouse
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
WWW
Fontes Externas
DB
DB
DB
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados


Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Data Marts
Data Warehouse
DB
Repositório de Metadados
Camada de Data Warehouse
Fontes de Dados
contêm dados operacionais
Camada de Pré-processamento dos Dados
WWW
Fontes Externas
DB
DB
DB
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados


Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Fontes de Dados
Autônomas
Distribuídas
Heterogêneas
SGBDs relacionais, objeto-relacionais
Documentos HTML, SGML
Bases de conhecimento, bases NoSQL
Planilhas, arquivos
Camada de Pre-processamento dos Dados
www
DB
DB
DB
Fontes Externas
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados

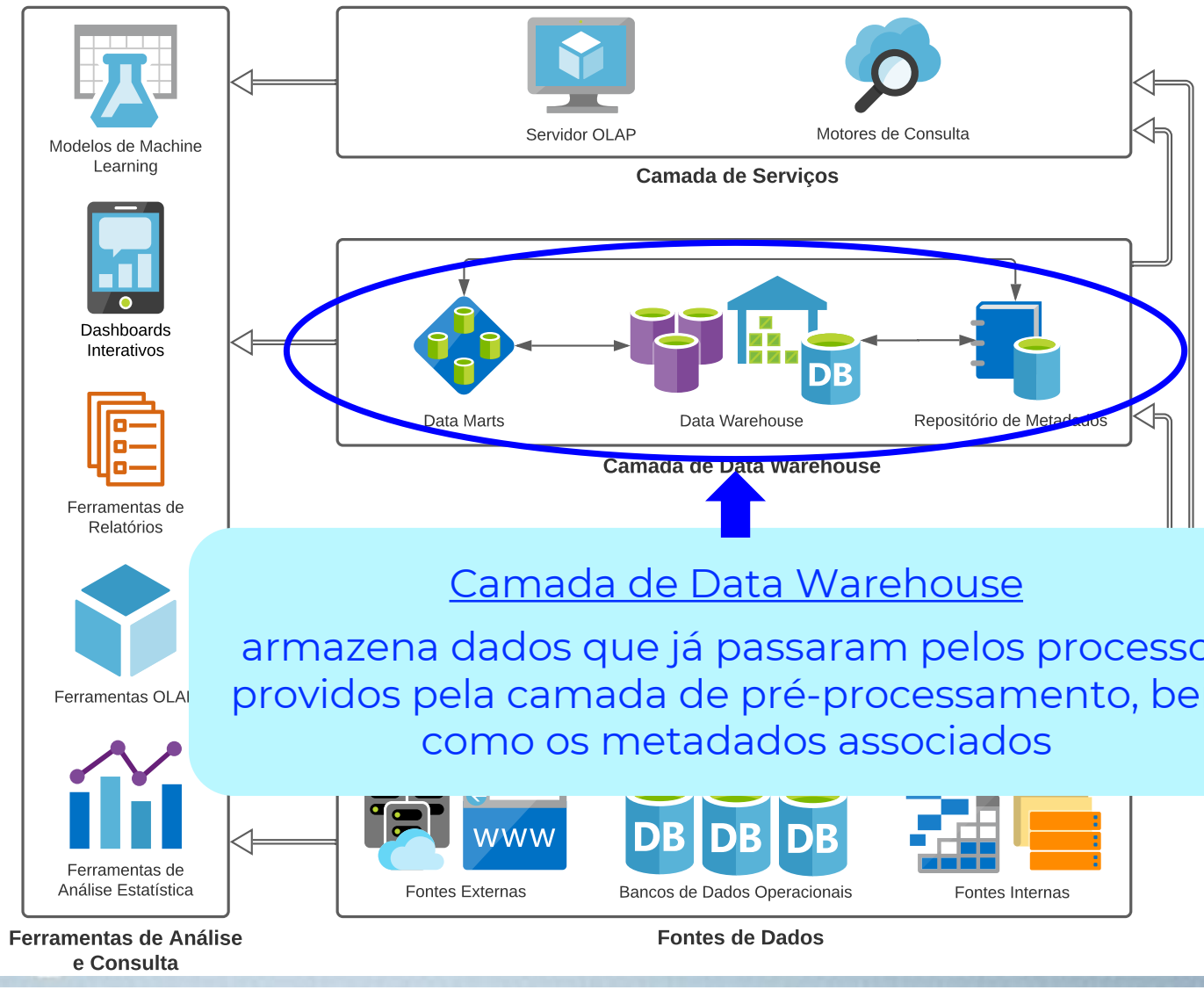


Camada de Pré-processamento dos Dados realiza a preparação dos dados


Modelos de Machine Learning
Servidor OLAP
Motores de Consulta
Camada de Serviços
Processos de ETL/ELT
Extração
Transformação
Carga
ETL
extração, transformação e carga
ELT
extração e carga no data lake
transformação e carga no DW
Ferramentas de Relatórios
Ferramentas OLAP
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
Ferramentas de Análise Estatística
WWW
DB
DB
DB
Fontes Externas
Bancos de Dados Operacionais
Fontes Internas
Ferramentas de Análise e Consulta
Fontes de Dados

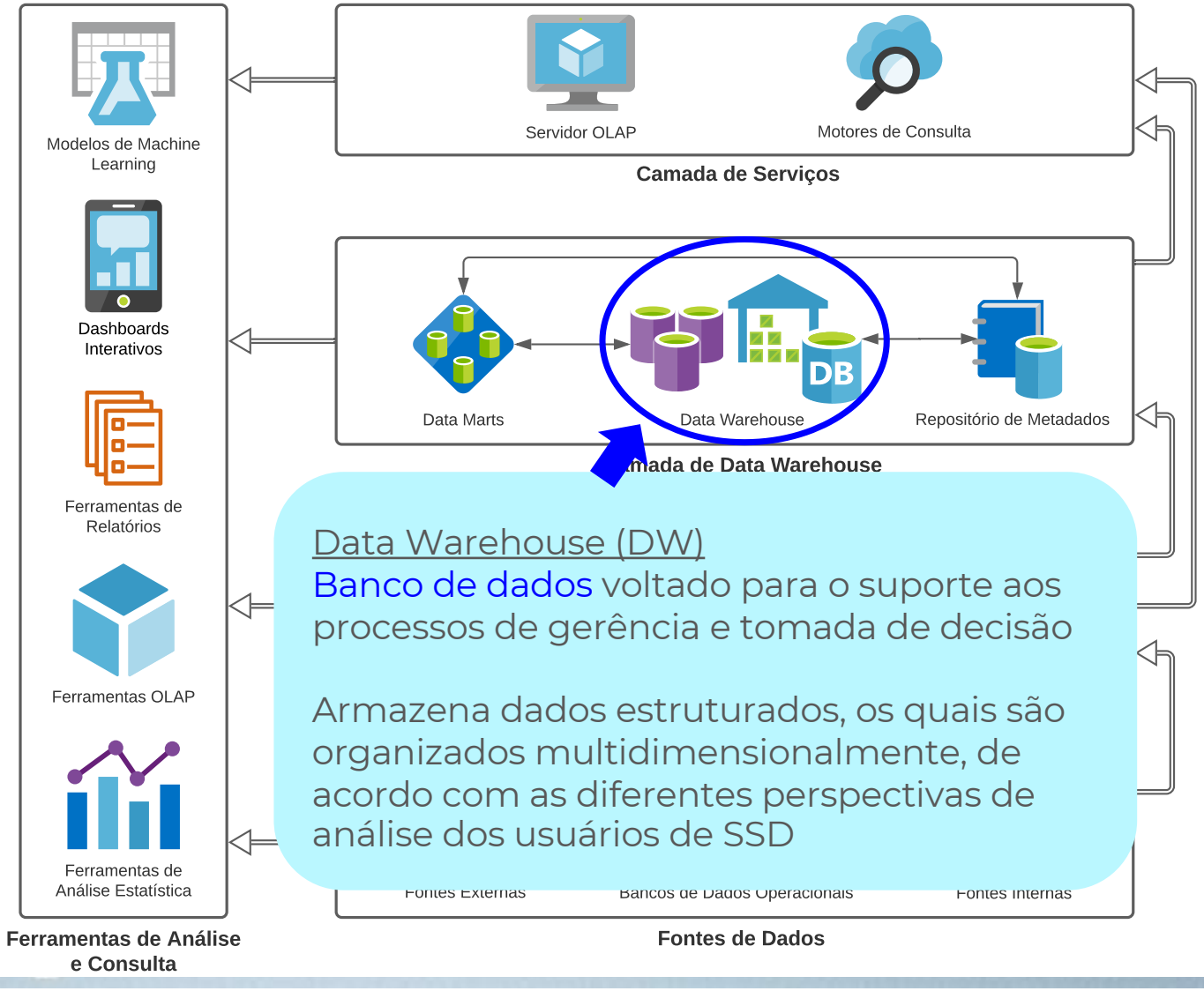

Data Lake
Área de armazenamento que contém um grande volume de dados estruturados, semiestruturados e não estruturados e que são processados somente quando a informação precisa ser obtida
tores de Consulta
Repositório de Metadados
Camada de Data Warehouse
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
WWW
Fontes Externas
DB
DB
DB
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados


Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Data Marts
Camada de Data Warehouse
Data Staging Area
Área de armazenamento que contém os dados extraídos das fontes, os quais vão passando por sucessivas modificações até que estejam prontos para serem carregados no DW
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
WWW
Fontes Externas
DB DB DB
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados



Camada de Data Warehouse

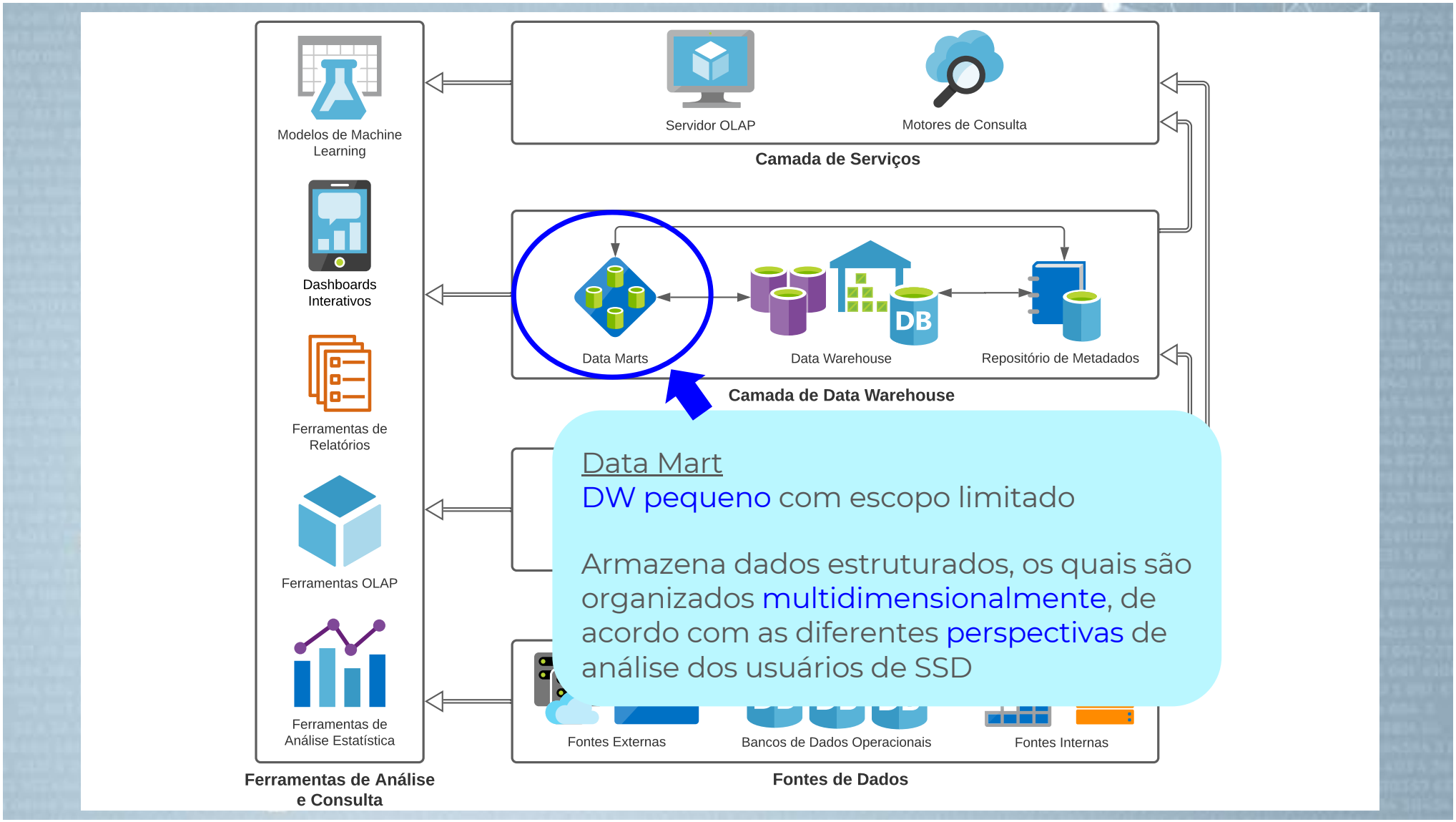
armazena dados que já passaram pelos processos providos pela camada de pré-processamento, bem como os metadados associados

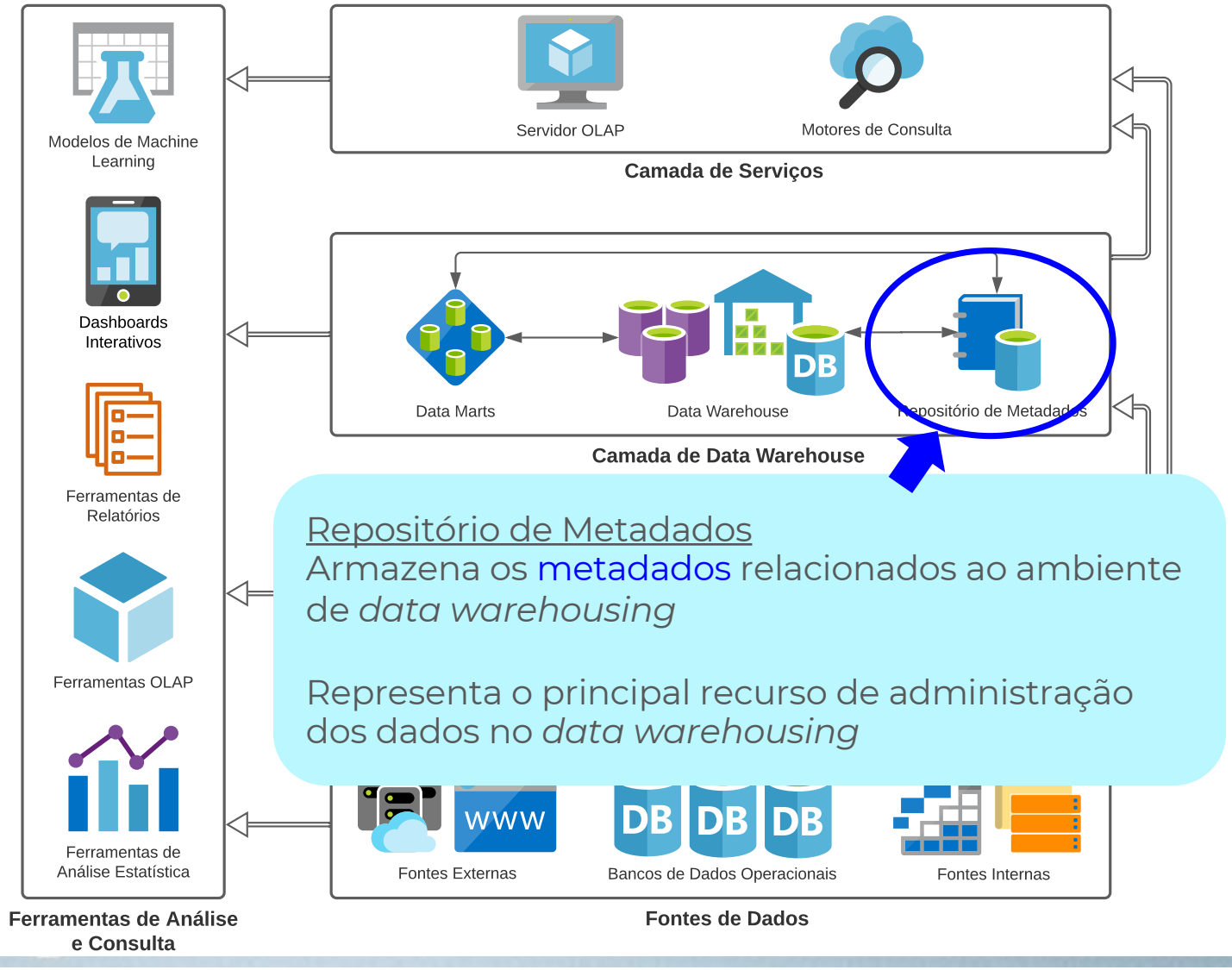


Data Warehouse (DW)

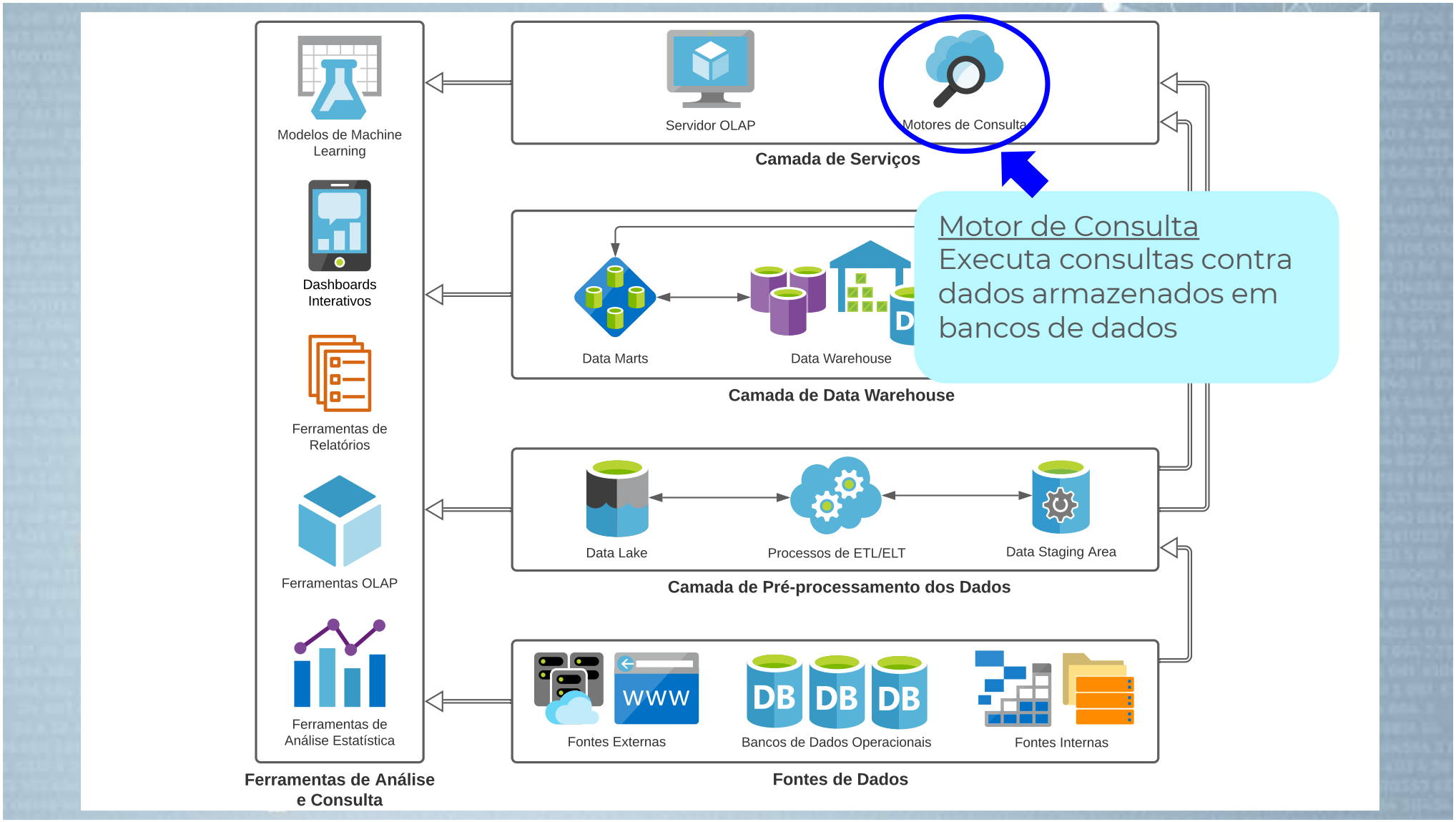
Banco de dados voltado para o suporte aos processos de gerência e tomada de decisão

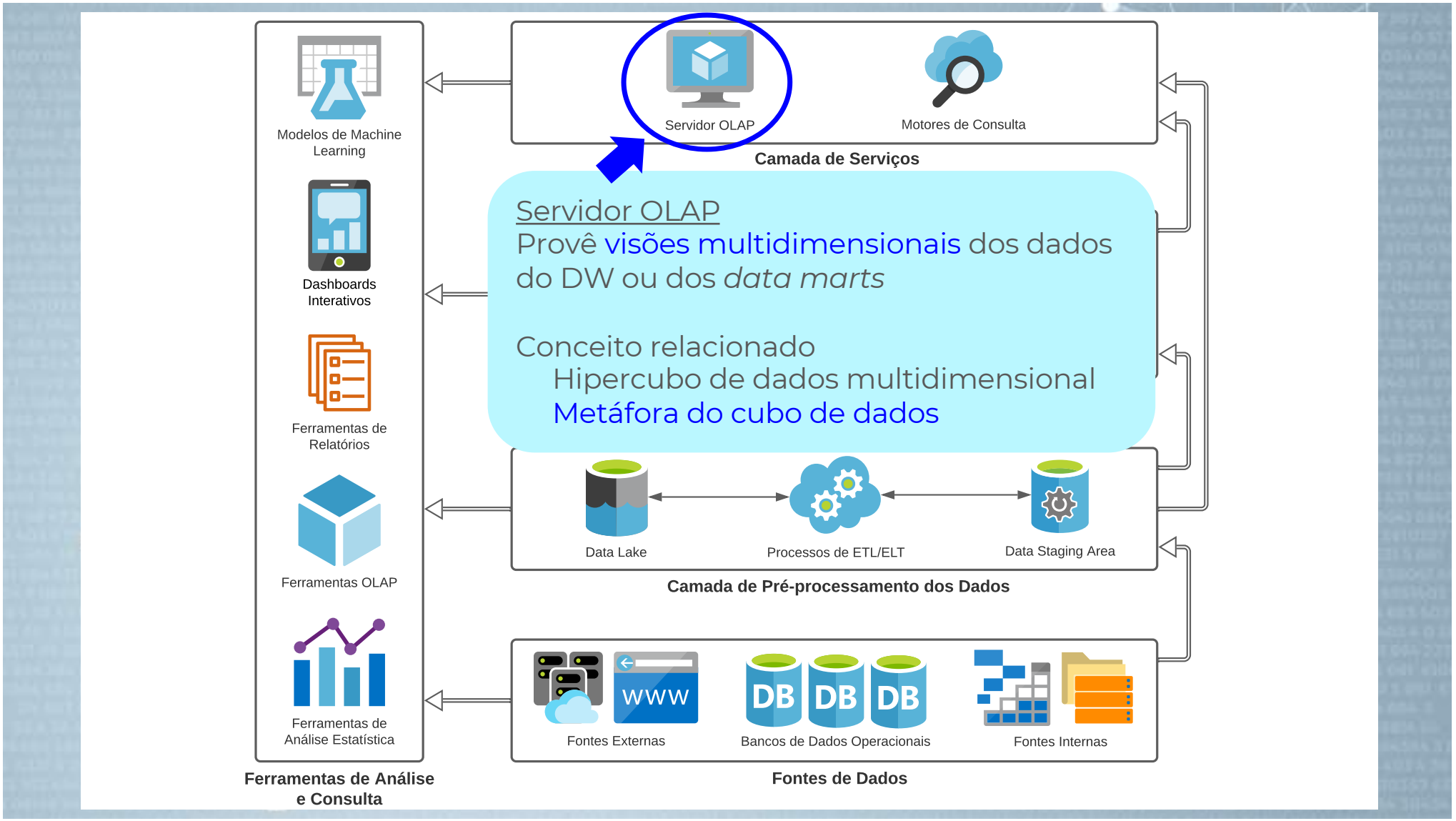
Armazena dados estruturados, os quais são organizados multidimensionalmente, de acordo com as diferentes perspectivas de análise dos usuários de SSD

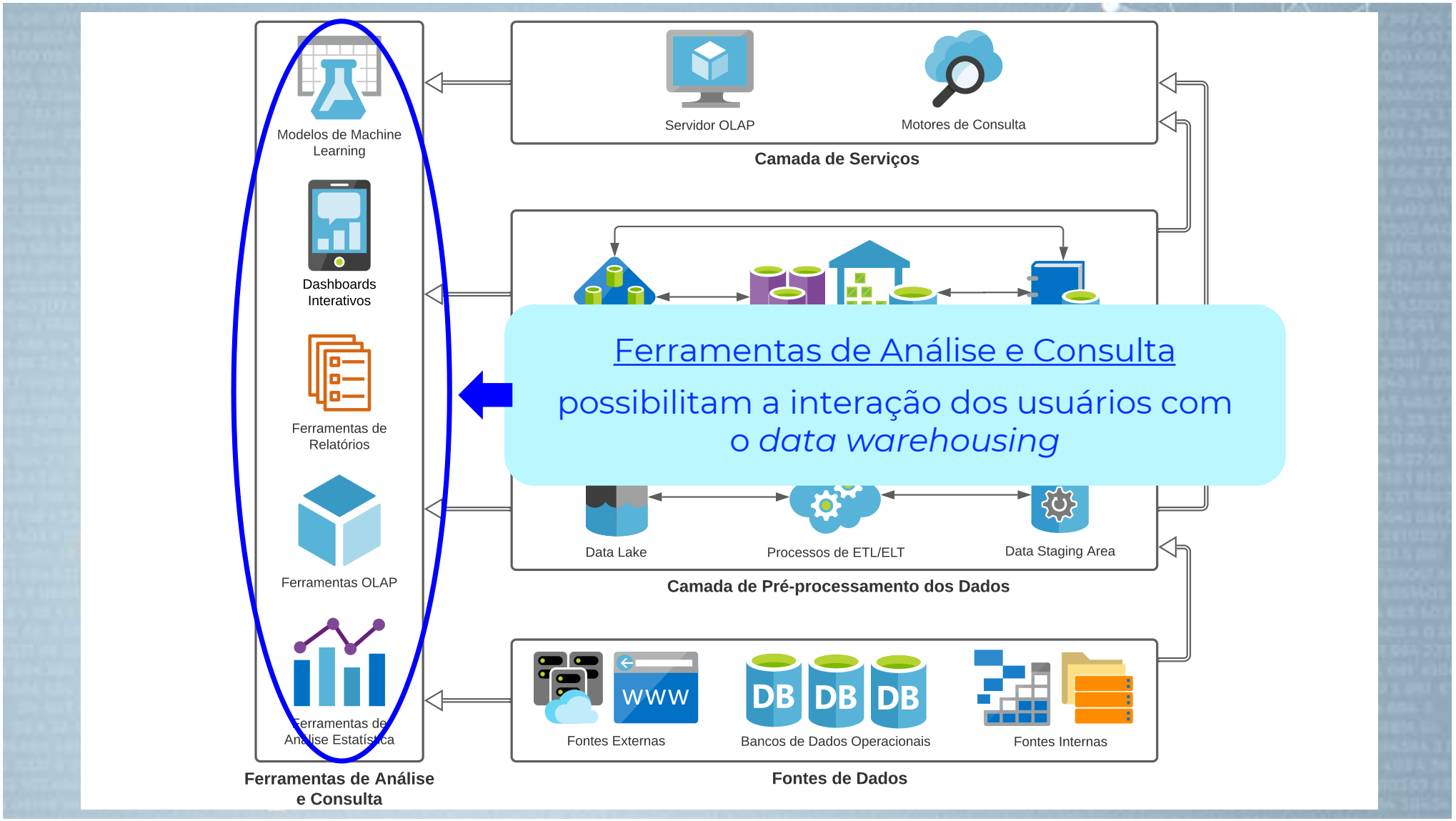

Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Data Marts
Data Warehouse
DB
Repositório de Metadados
Camada de Data Warehouse
Data Mart
DW pequeno com escopo limitado
Armazena dados estruturados, os quais são organizados multidimensionalmente, de acordo com as diferentes perspectivas de análise dos usuários de SSD
Fontes Externas
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados

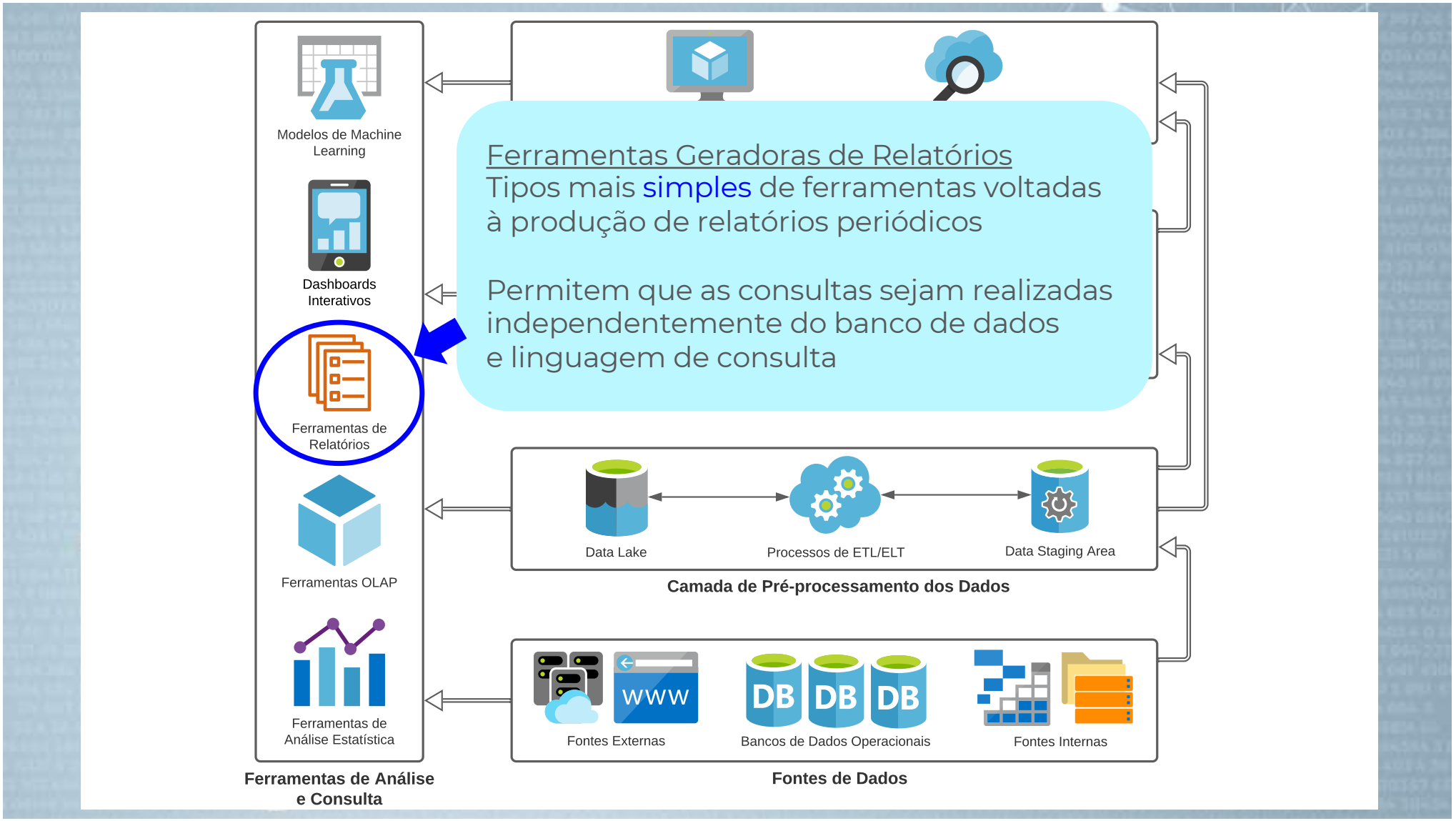

Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Data Marts
Data Warehouse
DB
Repositório de Metadados
Camada de Data Warehouse
Repositório de Metadados
Armazena os metadados relacionados ao ambiente de data warehousing
Representa o principal recurso de administração dos dados no data warehousing
www
Fontes Externas
DB DB DB
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados


Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Camada de Serviços
oferece funcionalidades para facilitar o acesso aos dados
Camada de Data Warehouse
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
WWW
Fontes Externas
DB
DB
DB
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados

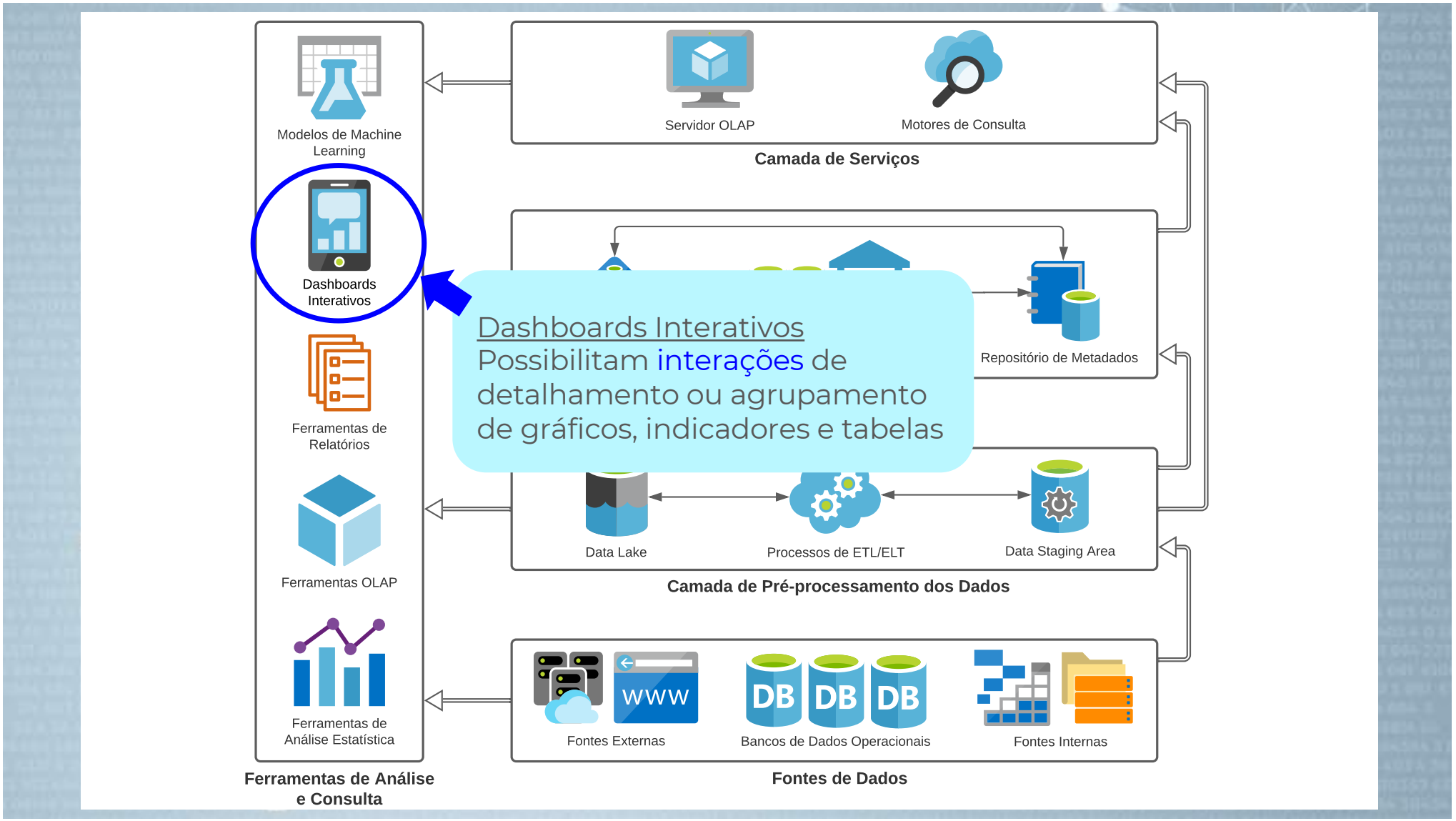

Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Motor de Consulta
Executa consultas contra dados armazenados em bancos de dados
Data Marts
Data Warehouse
Camada de Data Warehouse
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
WWW
Fontes Externas
DB
DB
DB
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados

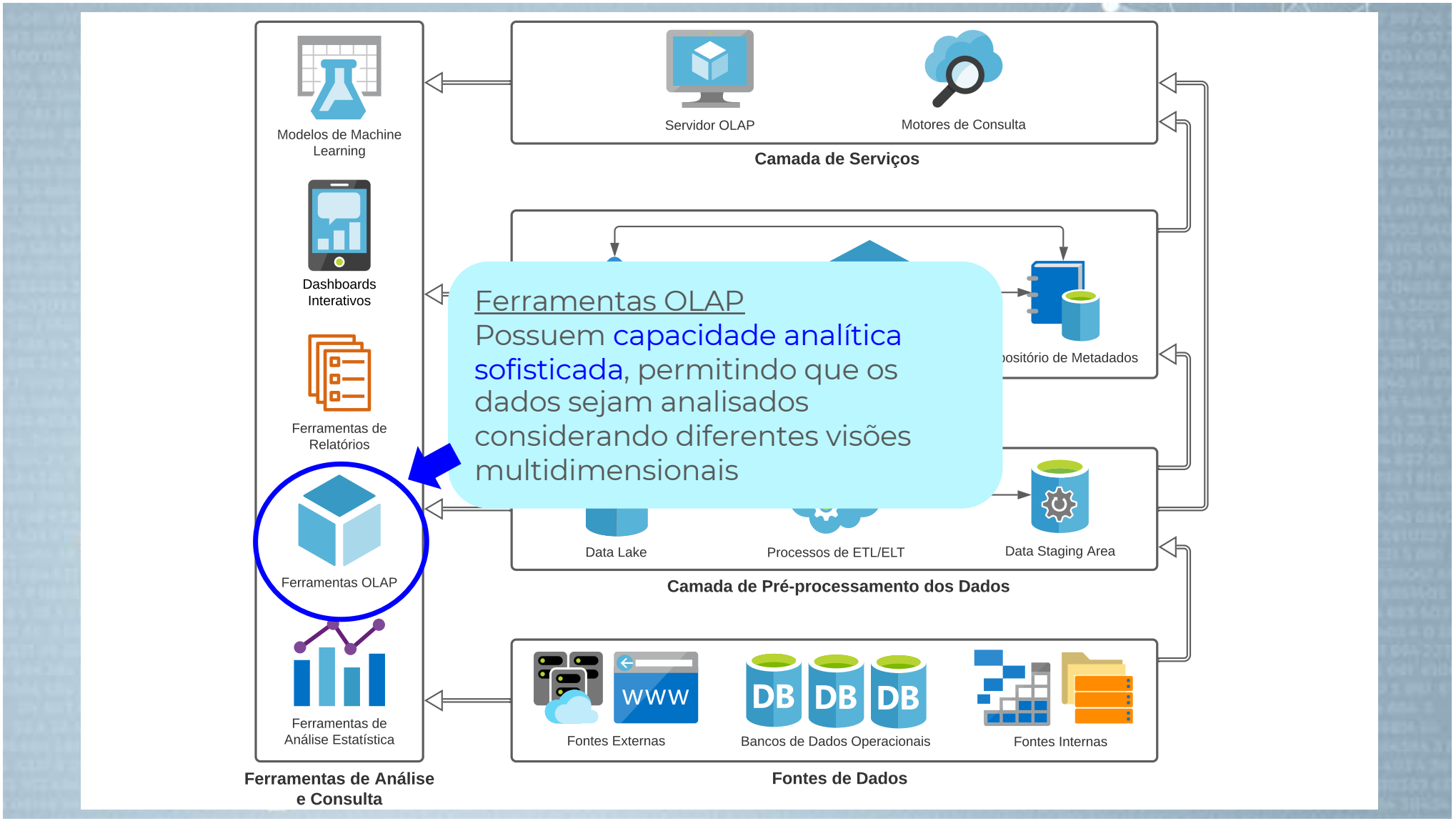

Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Servidor OLAP
Provê visões multidimensionais dos dados do DW ou dos data marts
Conceito relacionado
Hipercubo de dados multidimensional
Metáfora do cubo de dados
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
WWW
Fontes Externas
DB
DB
DB
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados

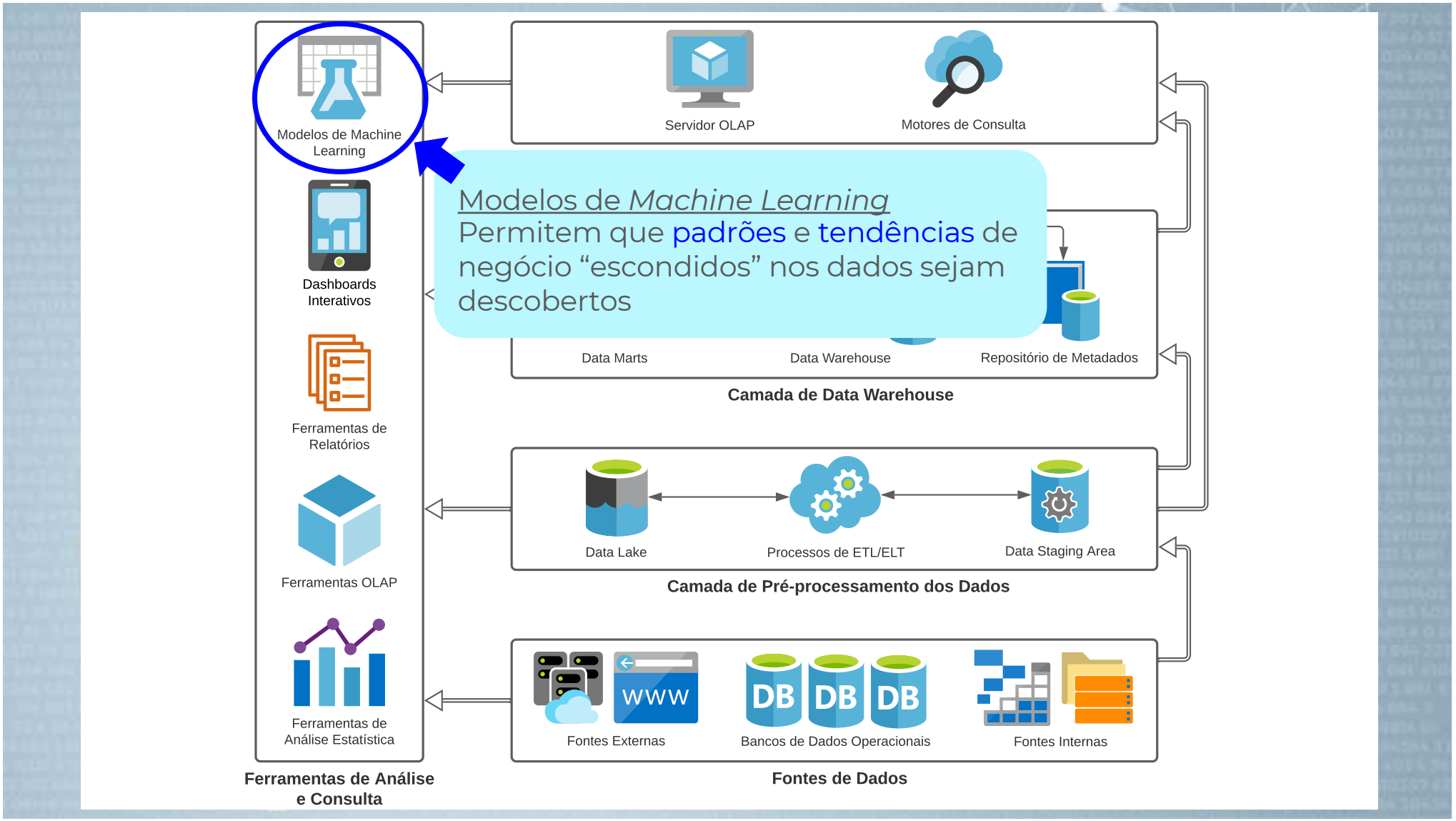

Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
Fontes Externas
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados
Ferramentas de Análise e Consulta
possibilitam a interação dos usuários com o data warehousing

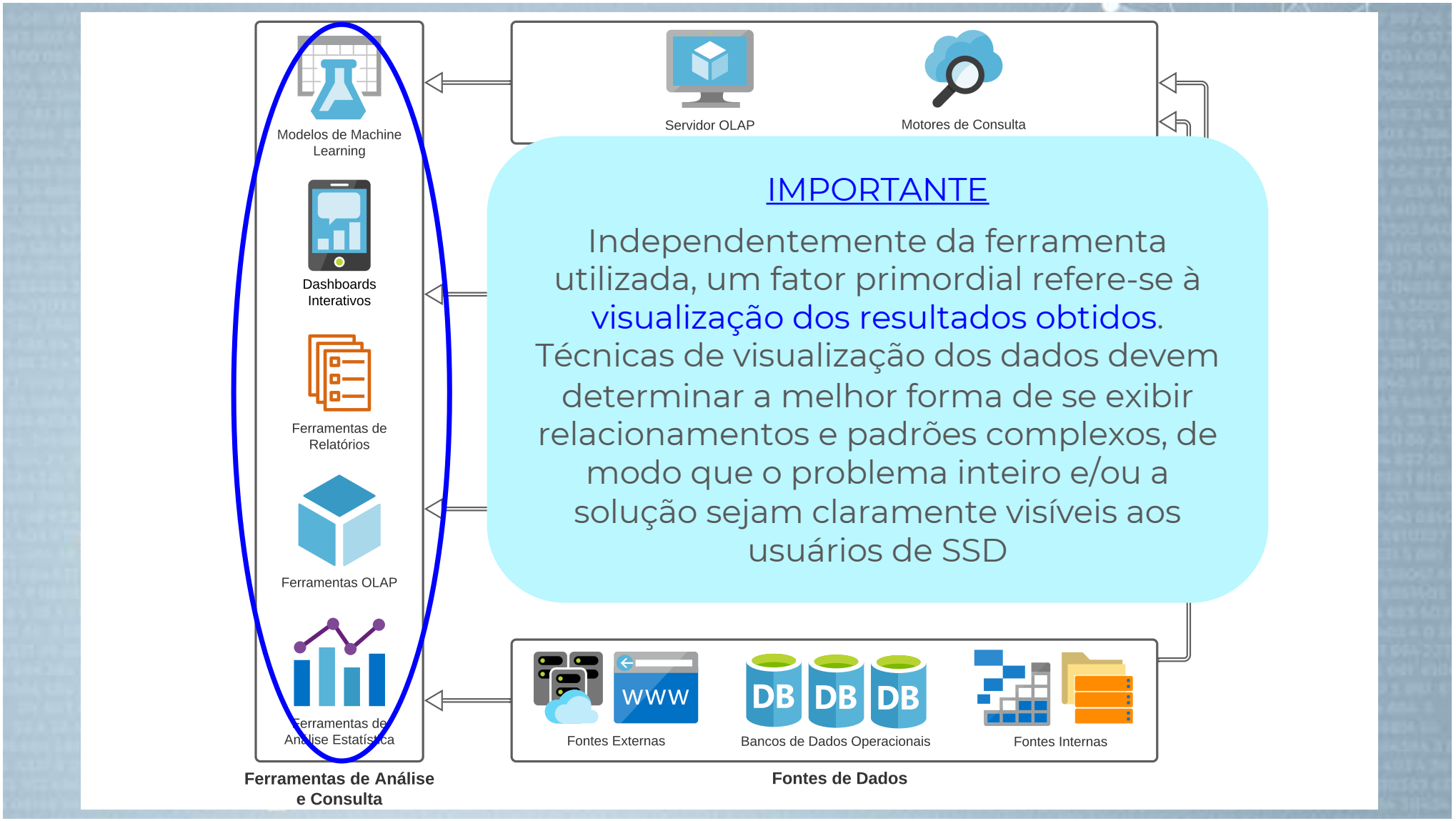

Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Ferramentas Geradoras de Relatórios
Tipos mais simples de ferramentas voltadas à produção de relatórios periódicos
Permitem que as consultas sejam realizadas independentemente do banco de dados e linguagem de consulta
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
WWW
DB
DB
DB
Fontes Externas
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados


Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Data Marts
Data Warehouse
Repositório de Metadados
Camada de Data Warehouse
Staging Area
Fontes Externas
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados
Ferramentas de Análise Estatística
Analisam os dados usando métodos estatísticos


Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Repositório de Metadados
Dashboards Interativos
Possibilitam interações de detalhamento ou agrupamento de gráficos, indicadores e tabelas
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
WWW
Fontes Externas
DB
DB
DB
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados

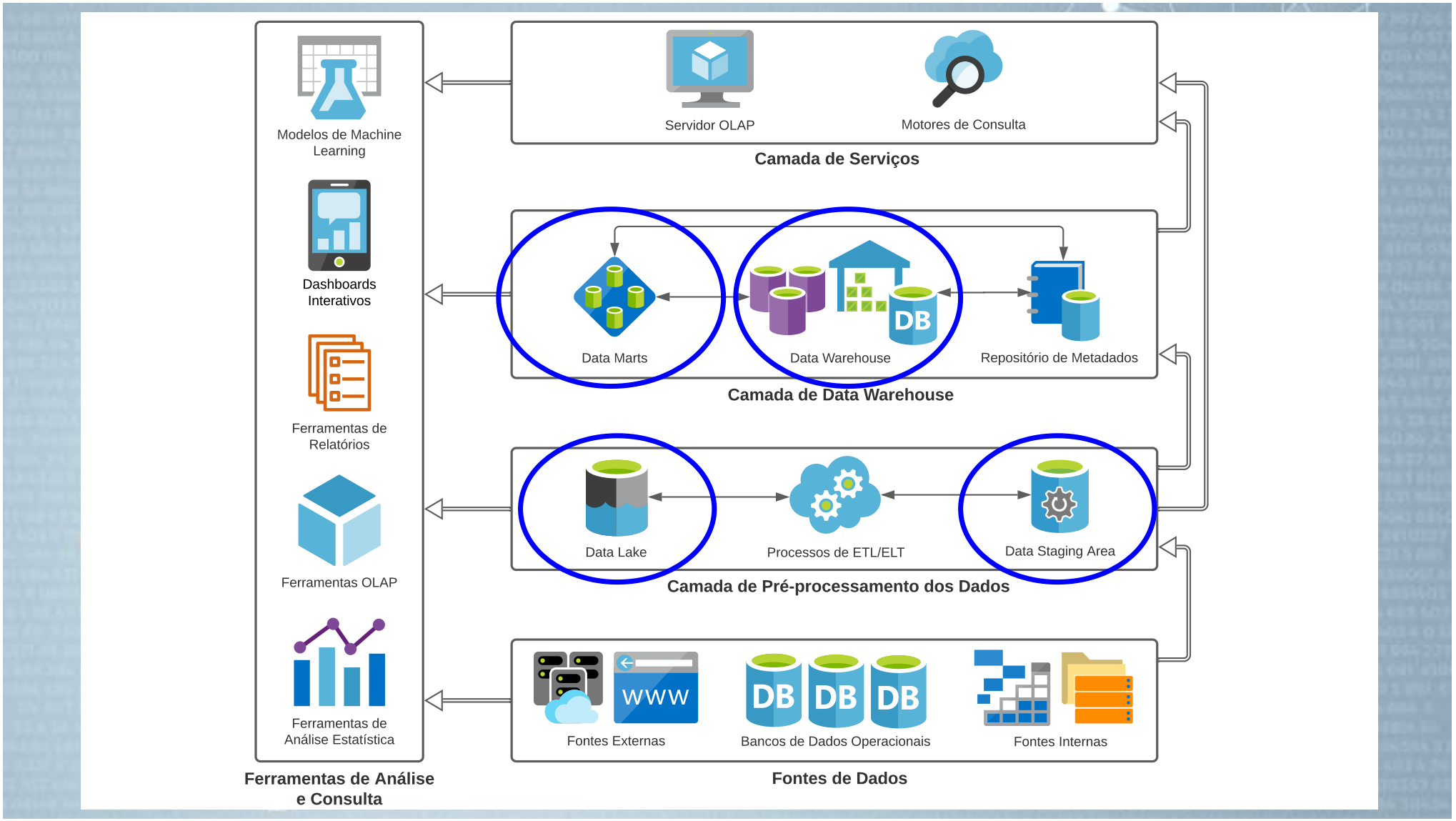

Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Repositório de Metadados
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
Fontes Externas
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados
Ferramentas OLAP
Possuem capacidade analítica sofisticada, permitindo que os dados sejam analisados considerando diferentes visões multidimensionais

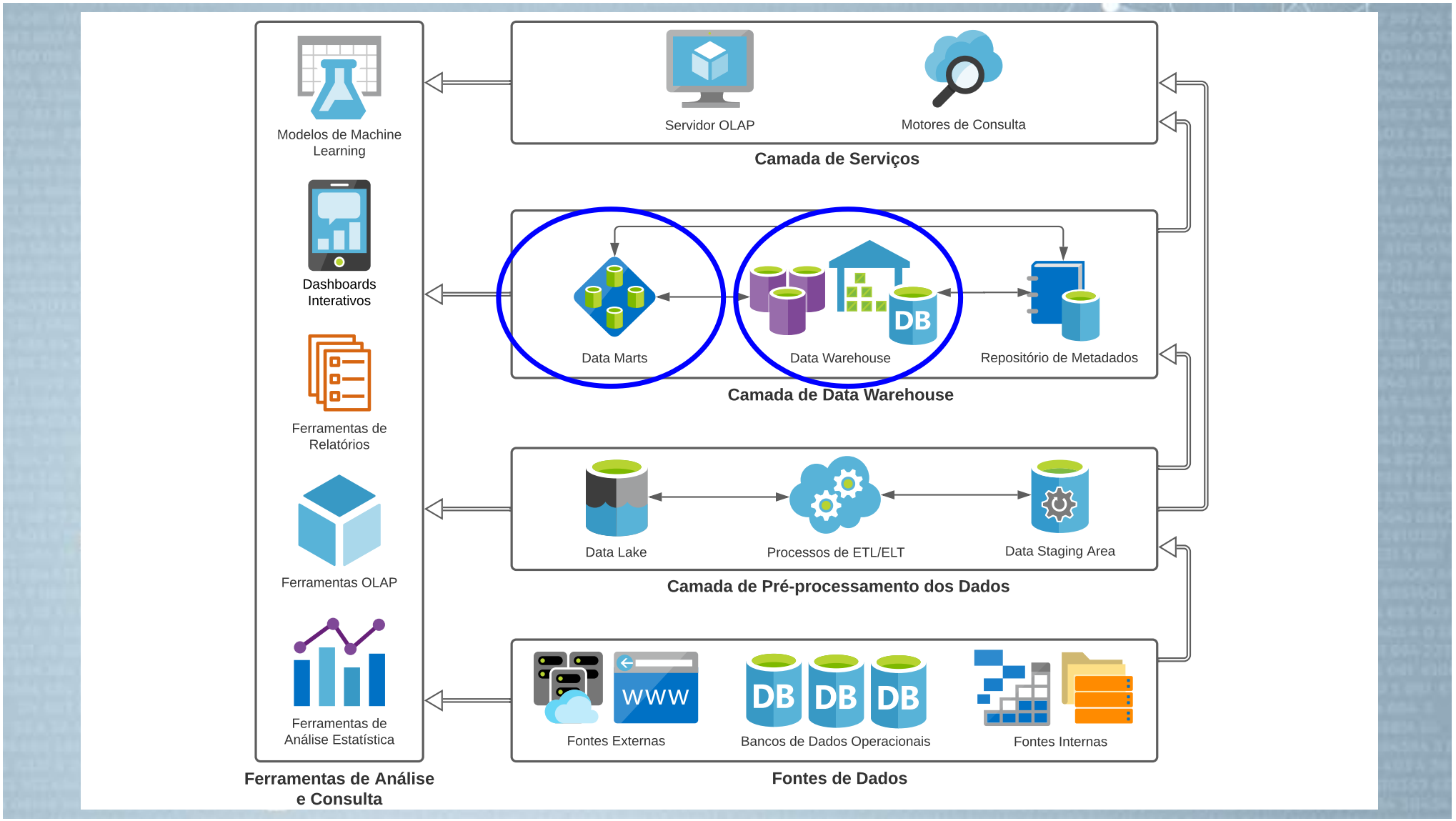

Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Data Marts
Data Warehouse
Repositório de Metadados
Camada de Data Warehouse
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
Fontes Externas
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados
Modelos de Machine Learning
Permitem que padrões e tendências de negócio "escondidos" nos dados sejam descobertos


Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
IMPORTANTE
Independentemente da ferramenta utilizada, um fator primordial refere-se à visualização dos resultados obtidos. Técnicas de visualização dos dados devem determinar a melhor forma de se exibir relacionamentos e padrões complexos, de modo que o problema inteiro e/ou a solução sejam claramente visíveis aos usuários de SSD
WWW
DB
DB
DB
Fontes Externas
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados

Agenda

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


Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Data Marts
Data Warehouse
DB
Repositório de Metadados
Camada de Data Warehouse
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
WWW
Fontes Externas
DB
DB
DB
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados


Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Data Marts
Data Warehouse
DB
Repositório de Metadados
Camada de Data Warehouse
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
WWW
Fontes Externas
DB
DB
DB
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados

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 material de consumo

Demanda: analisar gastos com salários de funcionários

Data Mart 1

Foco: salários e quantidadeLançamentos

Perspectivas: funcionário
cargo
filial
data

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

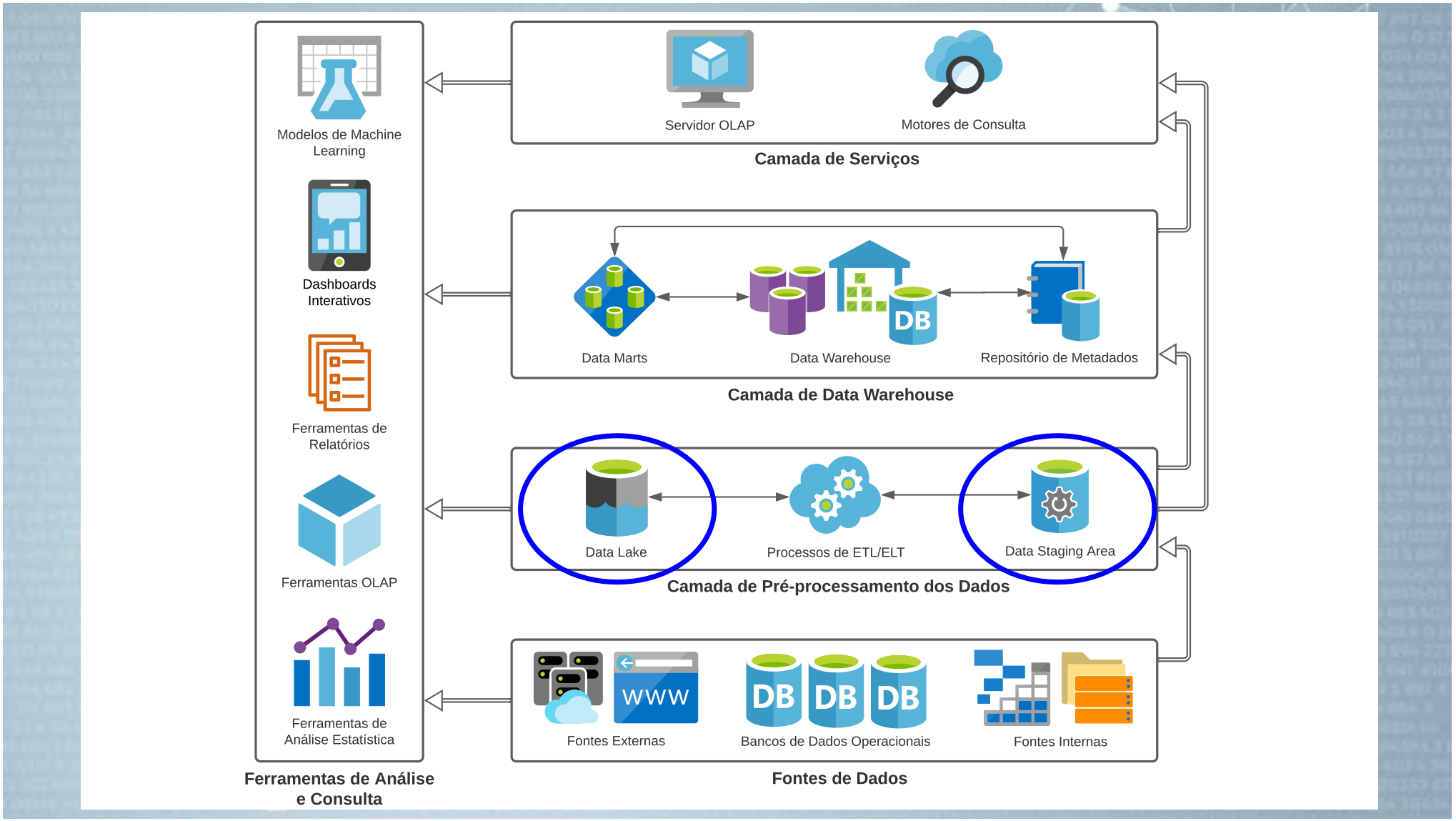
Foco: receitaTreinamento

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

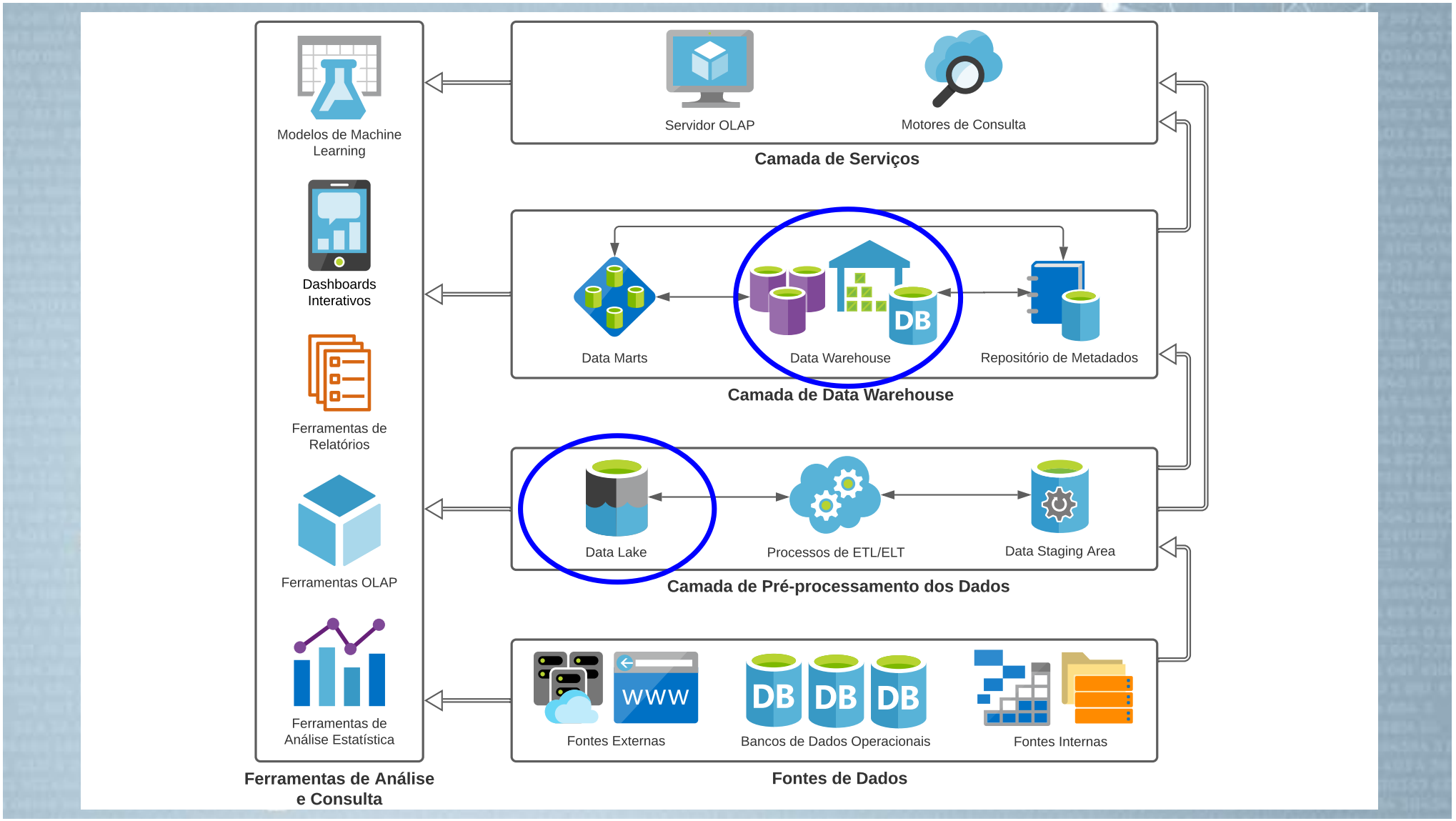

Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Data Marts
Data Warehouse
DB
Repositório de Metadados
Camada de Data Warehouse
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
WWW
Fontes Externas
DB
DB
DB
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados

Diferenças entre Data Staging Area e Data Lake

	Data Staging Area	Data Lake
Dados Armazenados	dados que sofrem modificações sucessivas	dados no formato nativo (*raw data*), 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	*data staging area* → DW	*data lake* → DW *data lake* → ferramentas de análise e consulta
Decorrência Histórica	processo ETL	processo ELT


Modelos de Machine Learning
Dashboards Interativos
Ferramentas de Relatórios
Ferramentas OLAP
Ferramentas de Análise Estatística
Ferramentas de Análise e Consulta
Servidor OLAP
Motores de Consulta
Camada de Serviços
Data Marts
Data Warehouse
DB
Repositório de Metadados
Camada de Data Warehouse
Data Lake
Processos de ETL/ELT
Data Staging Area
Camada de Pré-processamento dos Dados
WWW
Fontes Externas
DB
DB
DB
Bancos de Dados Operacionais
Fontes Internas
Fontes de Dados

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

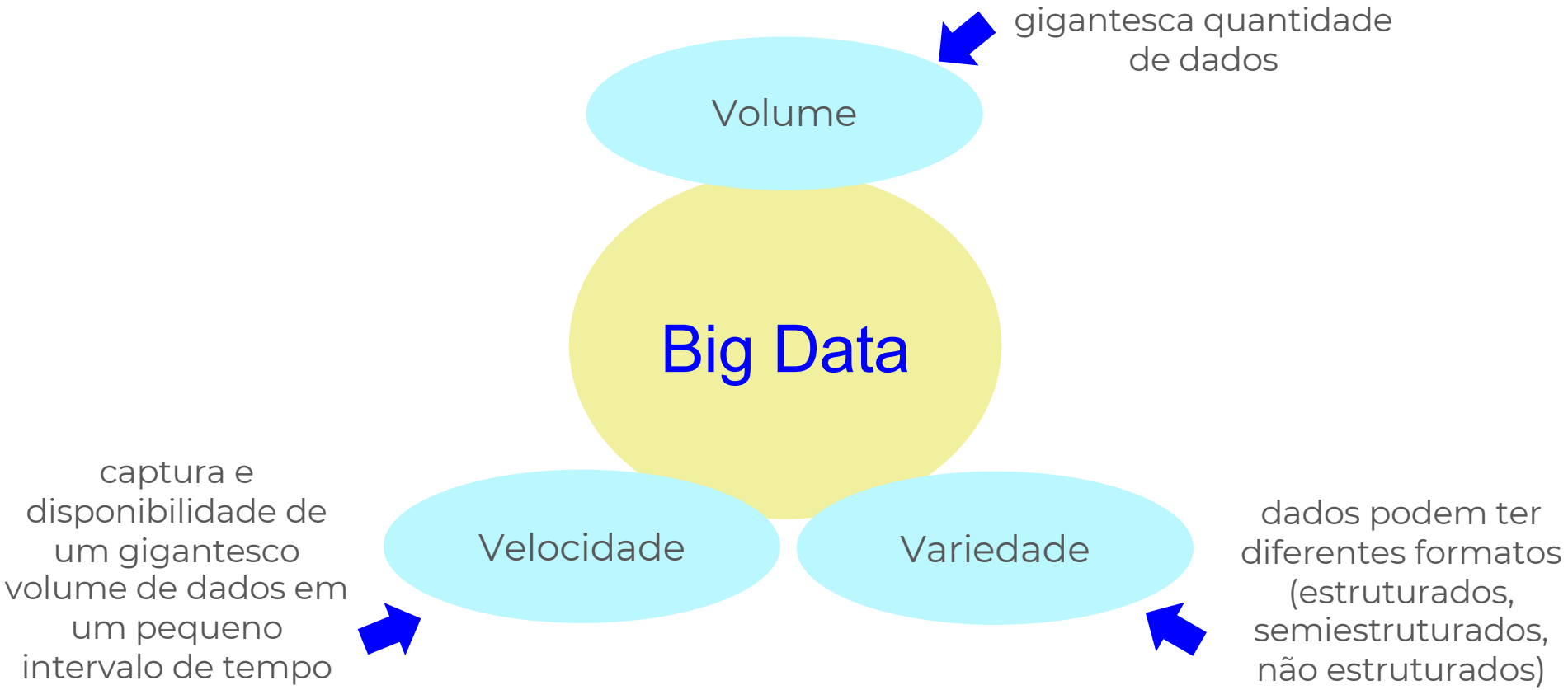
Diferenças entre Data Warehouse e Data Lake

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

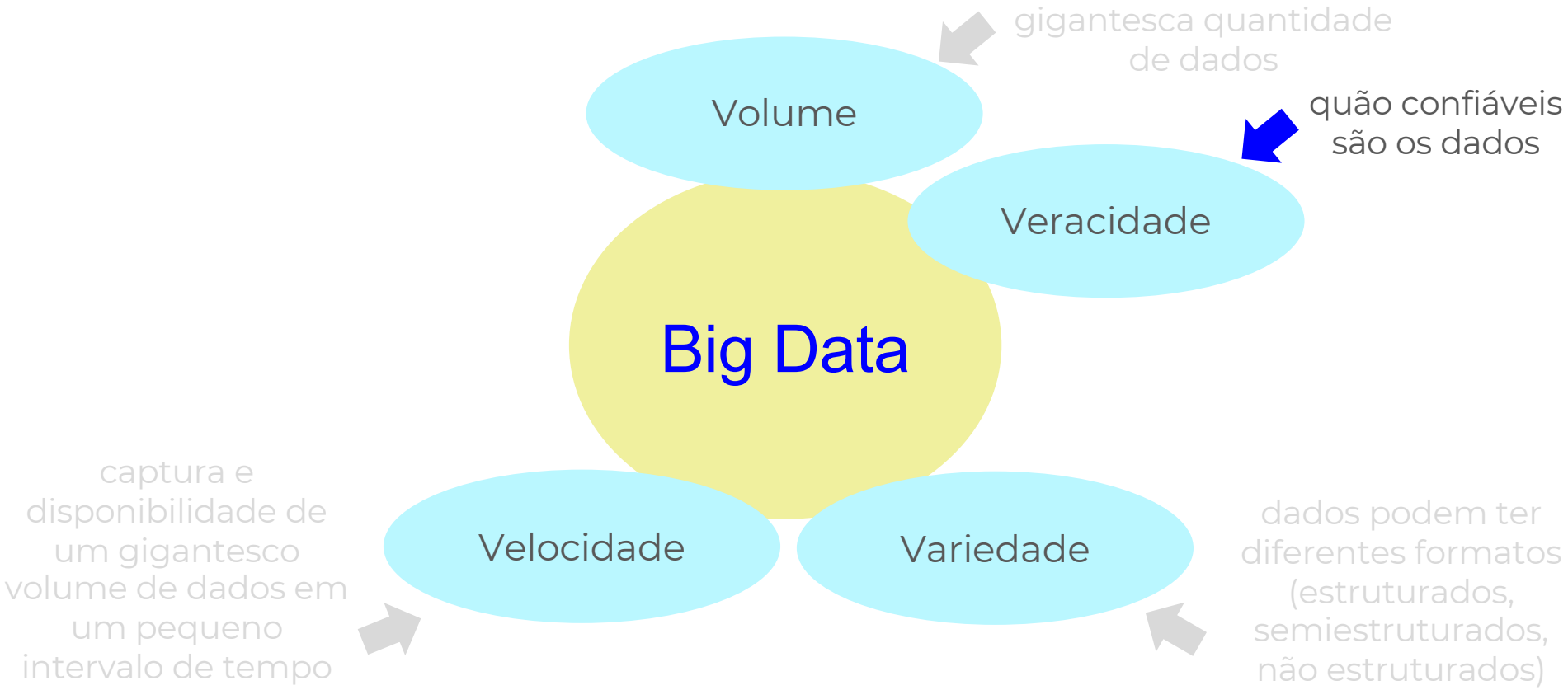
Agenda

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

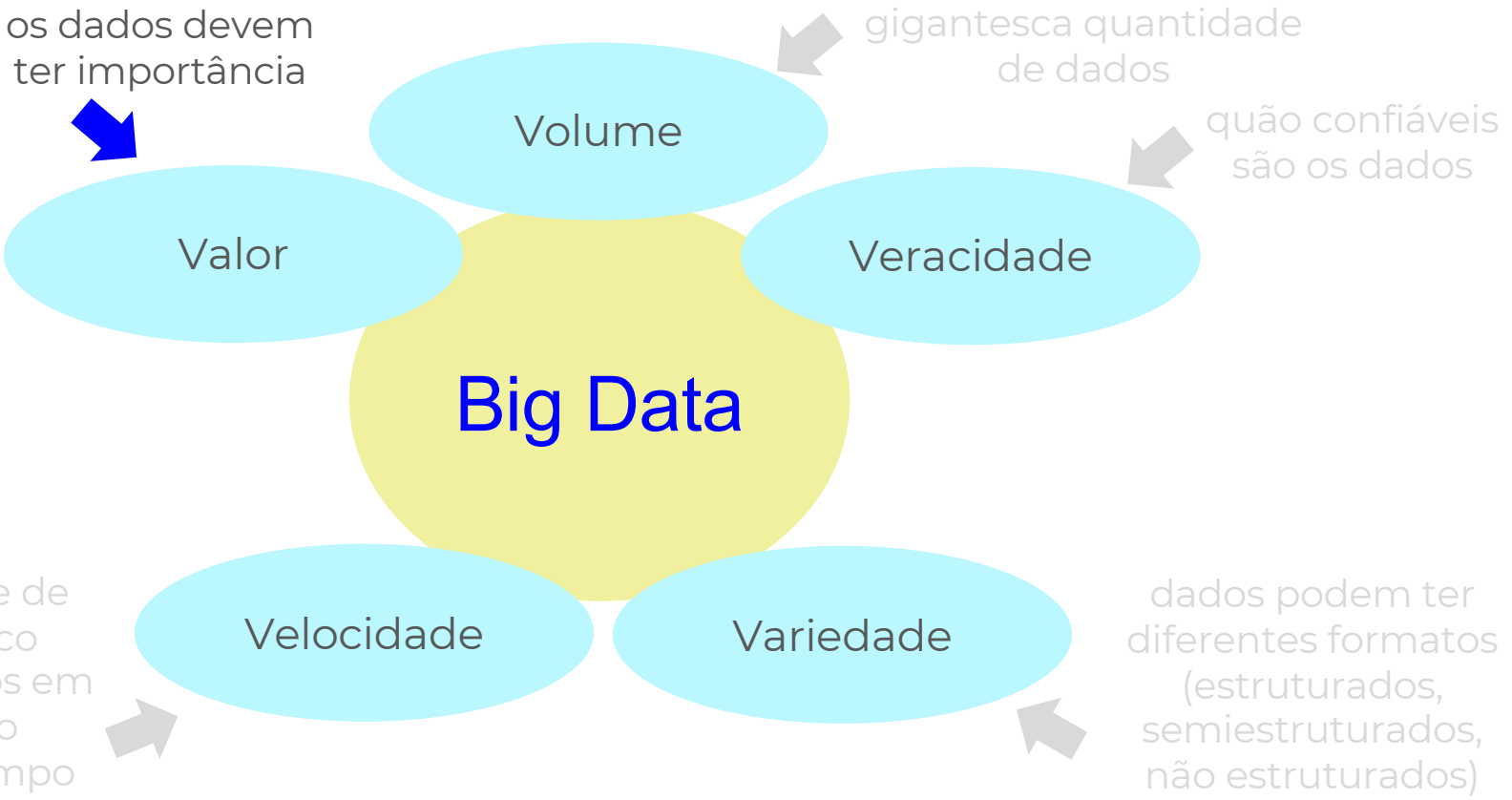
Modelo de 3Vs



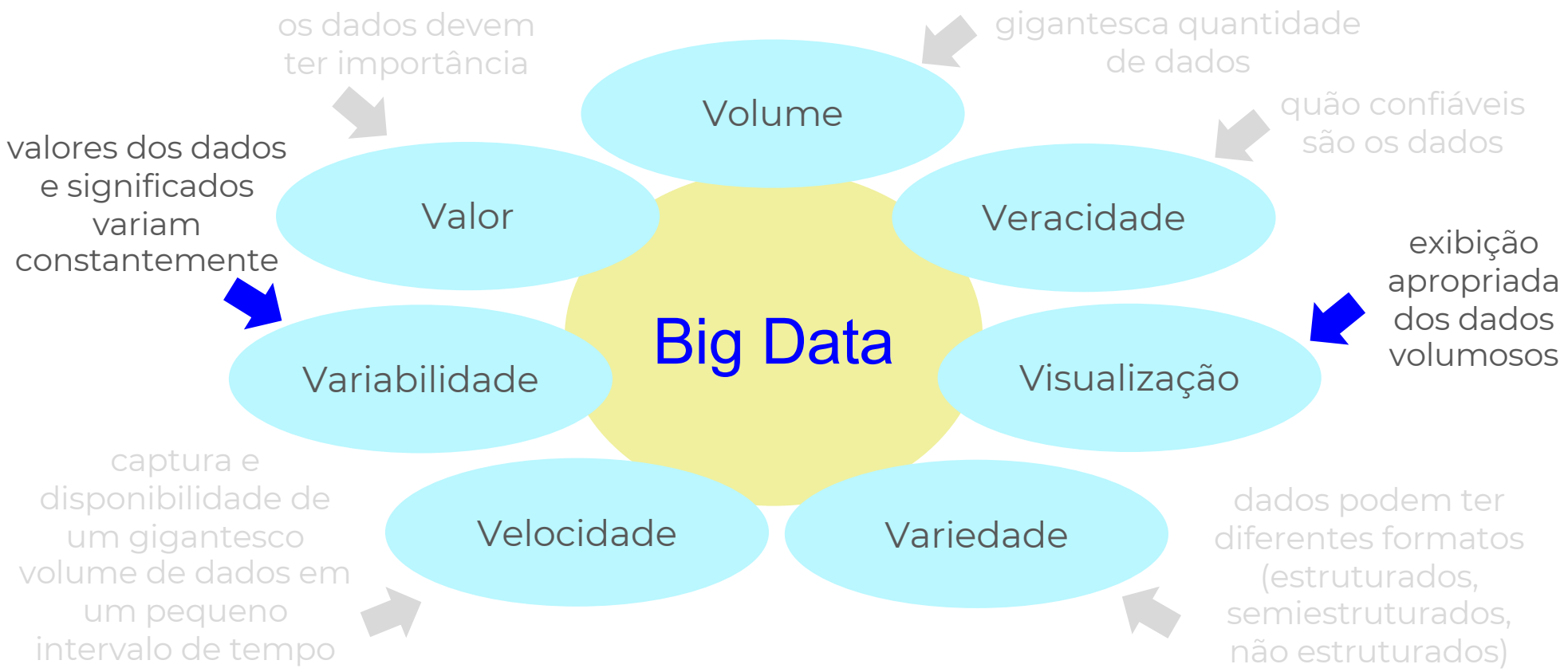
Modelo de 4Vs



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*)

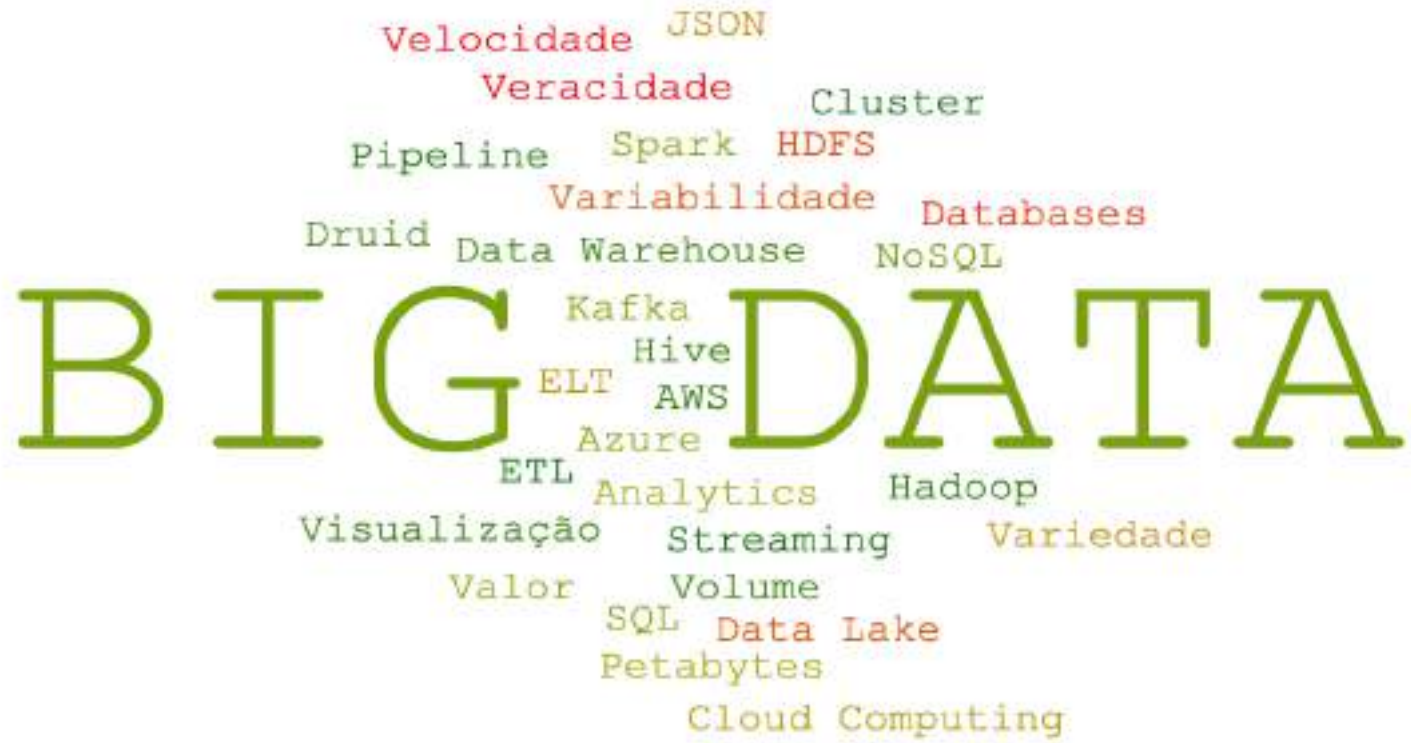
- Uso de *frameworks* de processamento paralelo e distribuído para simplificar a interação com os ambientes computacionais
  - Apache Hadoop
  - Apache Spark

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



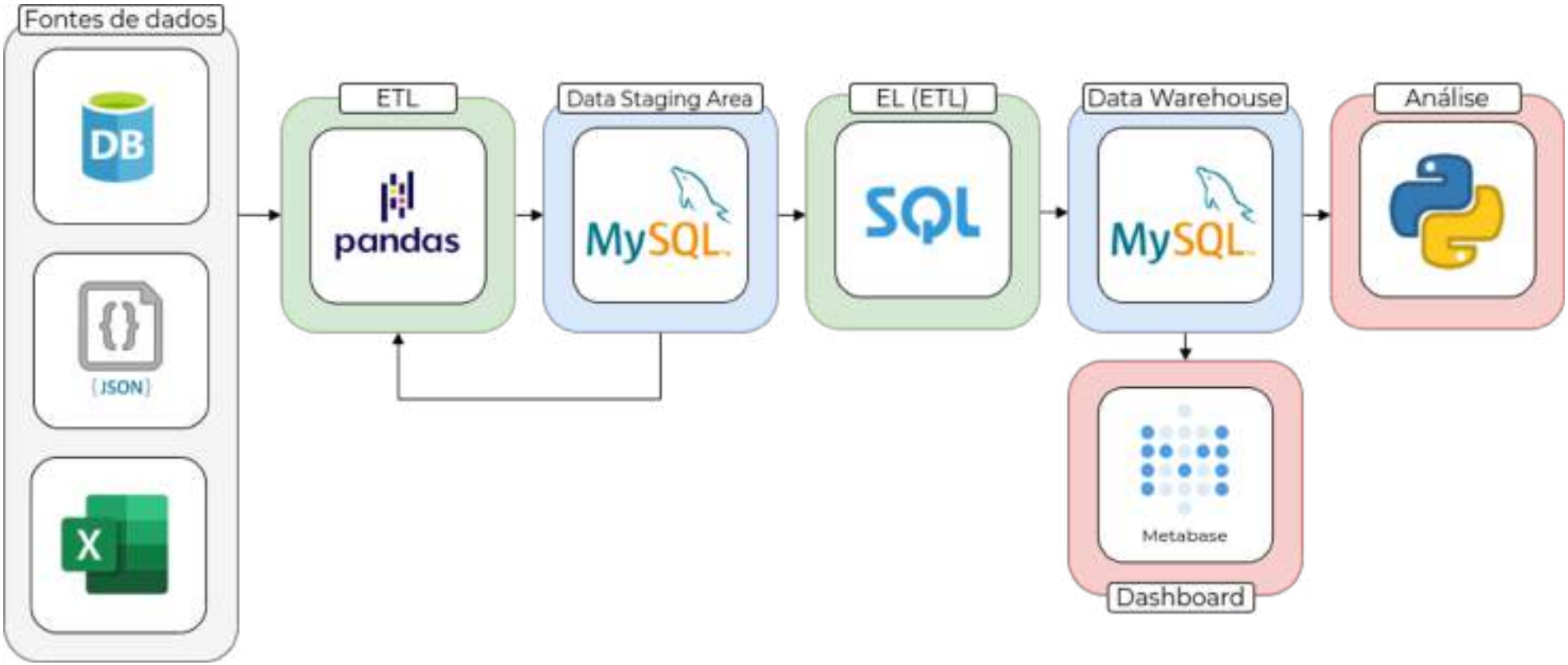
Agenda

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

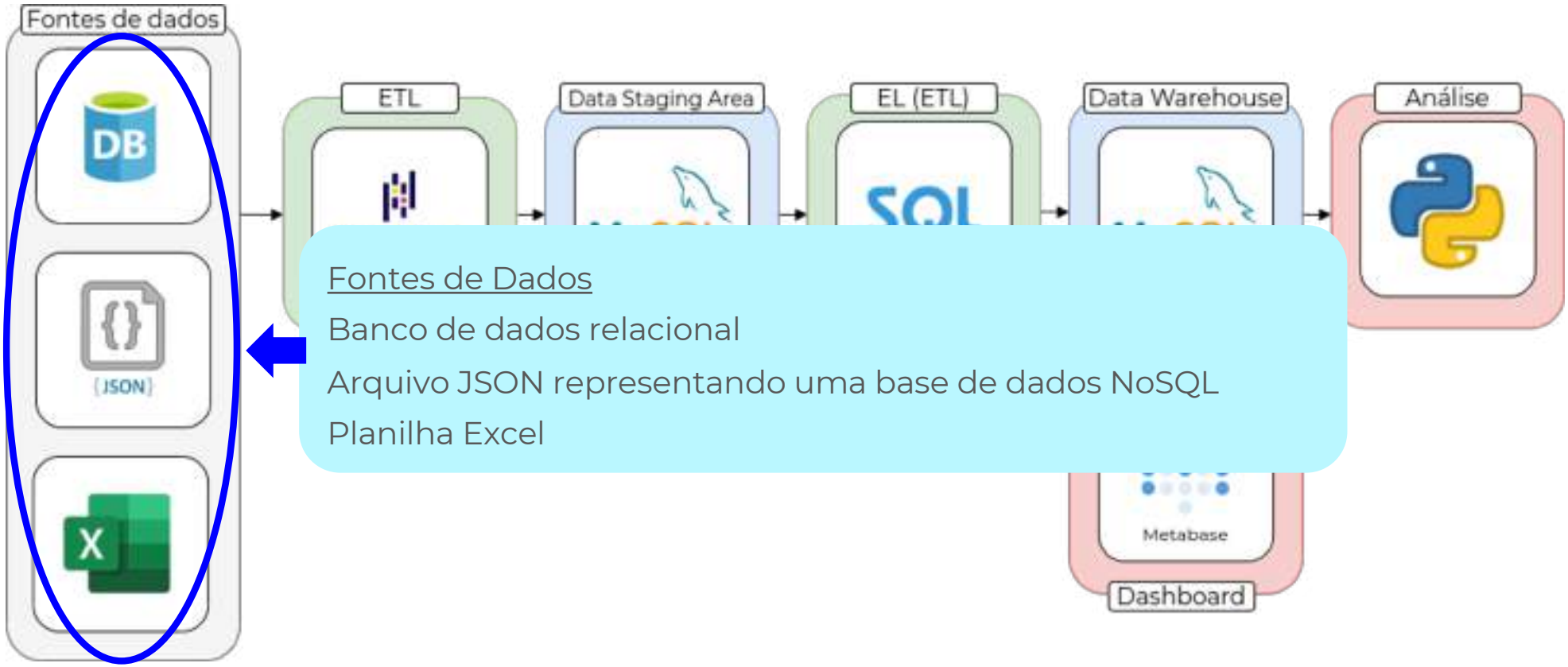
Exemplos de Pipeline

- Volumes de Dados Tradicionais
- Big Data
- Data Streaming

Processamento de Dados em Lote



Processamento de Dados em Lote



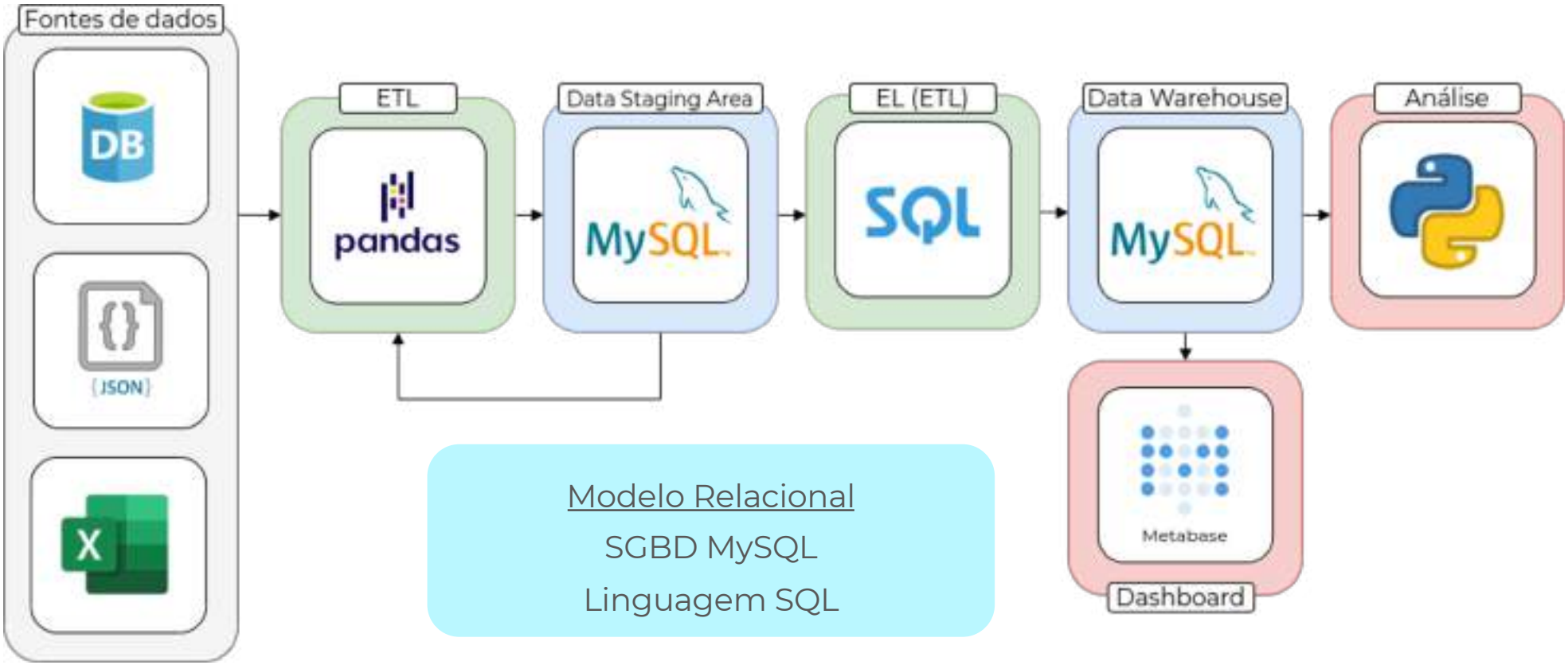
Fontes de Dados

Banco de dados relacional

Arquivo JSON representando uma base de dados NoSQL

Planilha Excel

Processamento de Dados em Lote

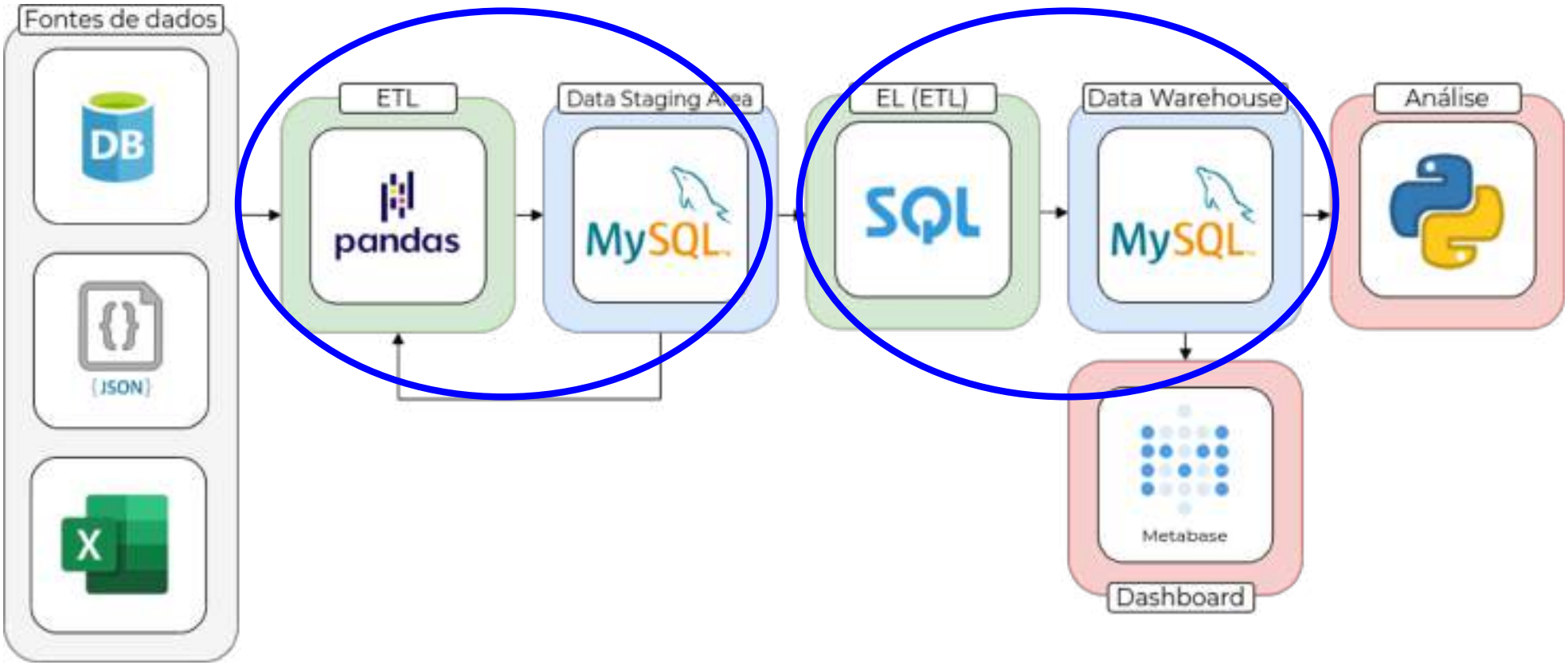


Modelo Relacional

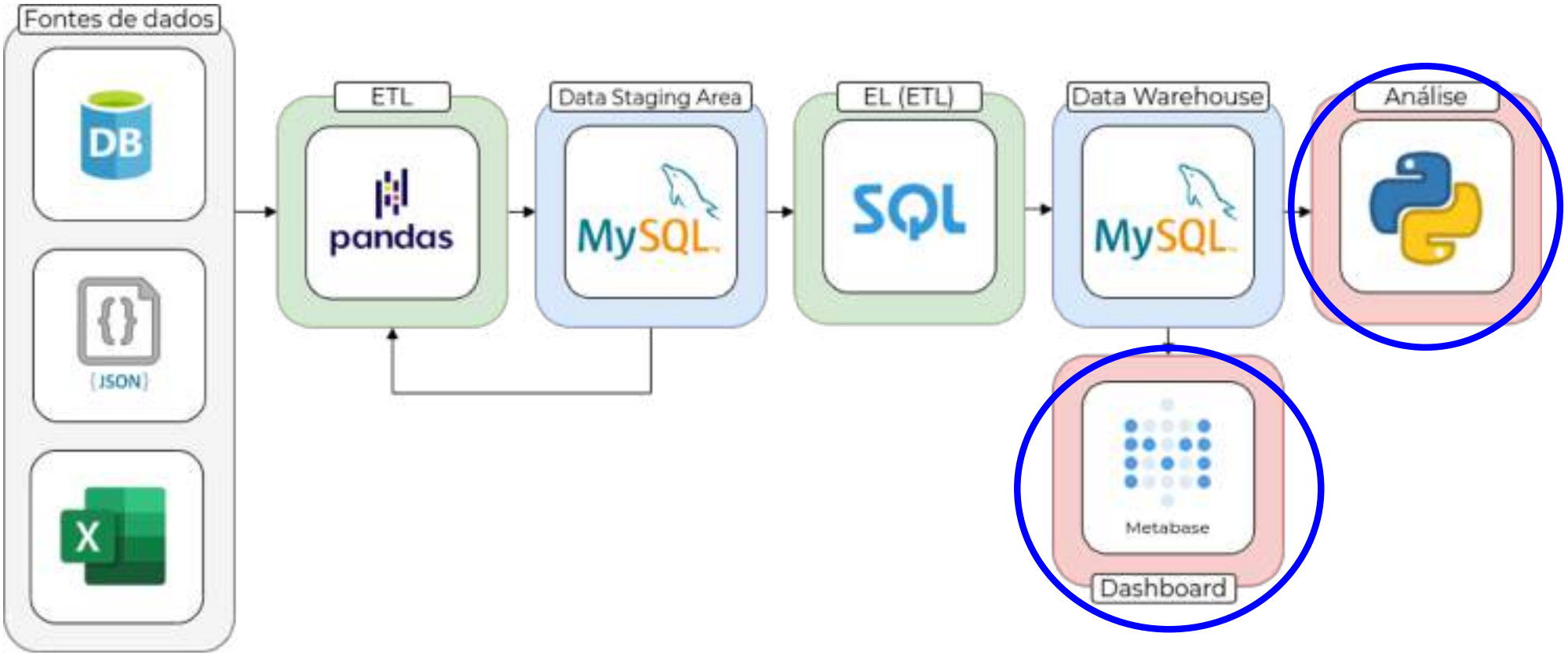
SGBD MySQL

Linguagem SQL

Processamento de Dados em Lote



Processamento de Dados em Lote

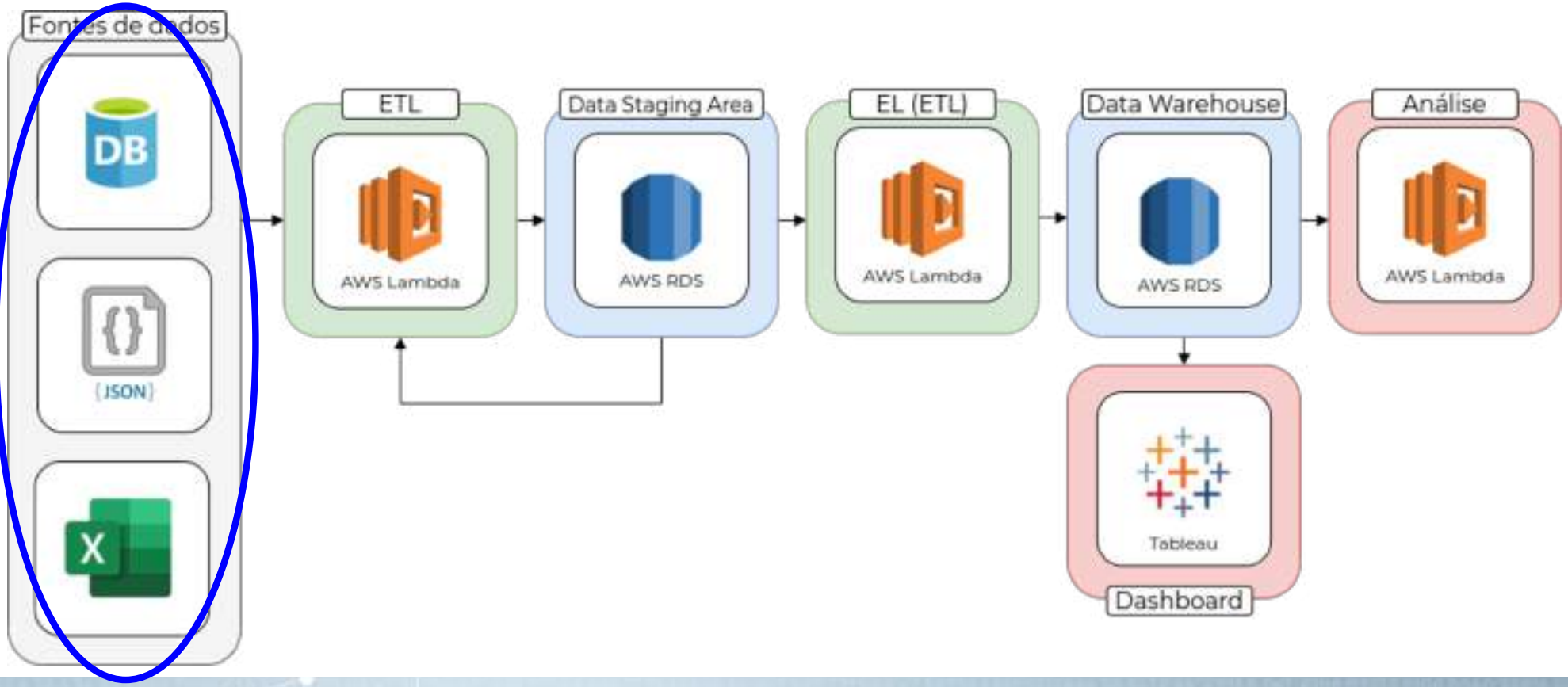


Processamento de Dados em Lote (Nuvem)

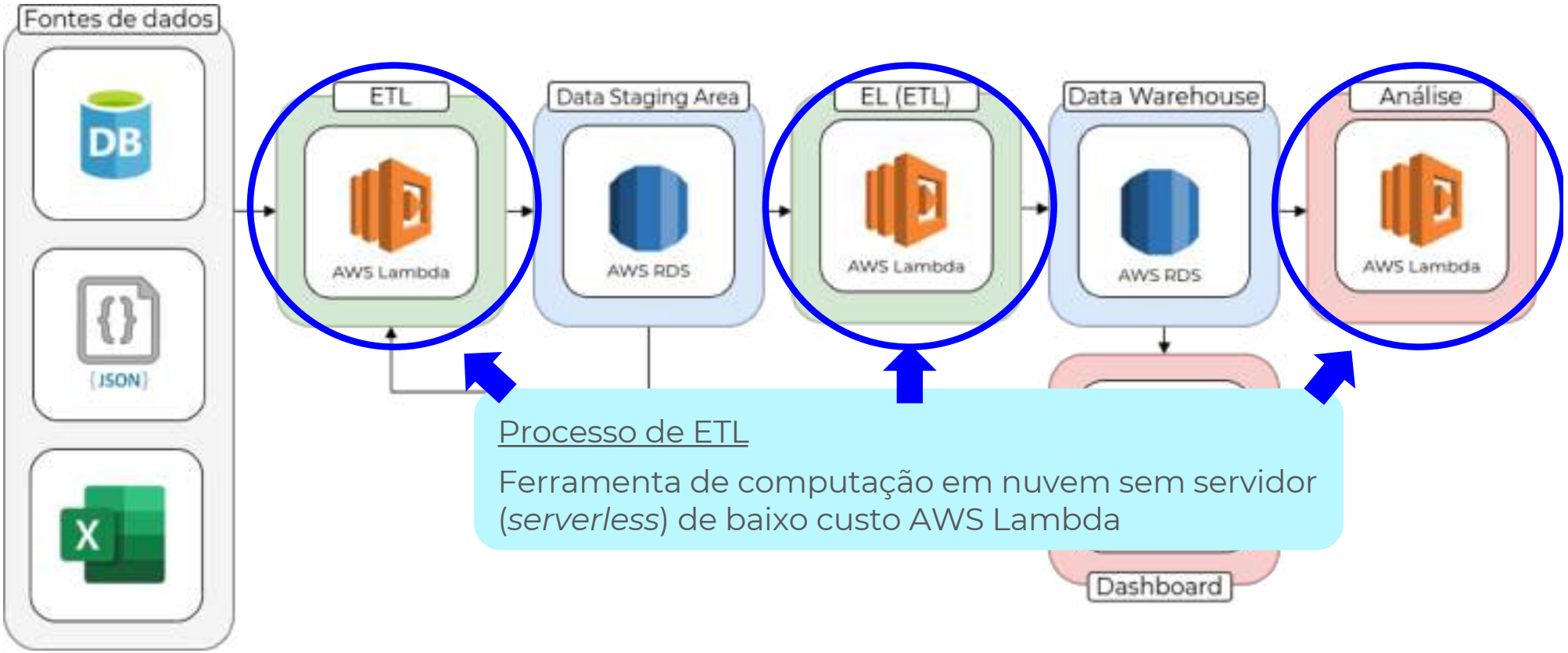
[Ambiente de Nuvem]

AWS (*Amazon Web Service*)

Processamento de Dados em Lote (Nuvem)



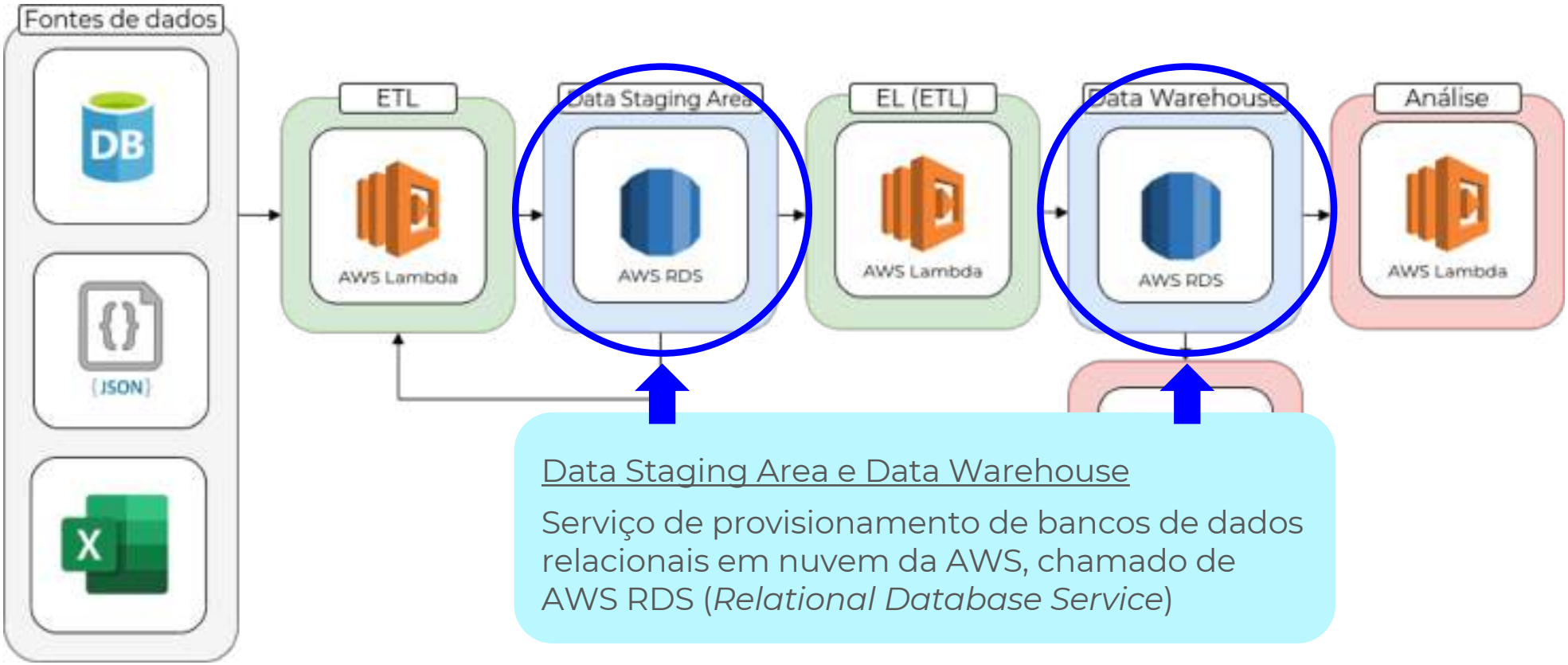
Processamento de Dados em Lote (Nuvem)



Processo de ETL

Ferramenta de computação em nuvem sem servidor (*serverless*) de baixo custo AWS Lambda

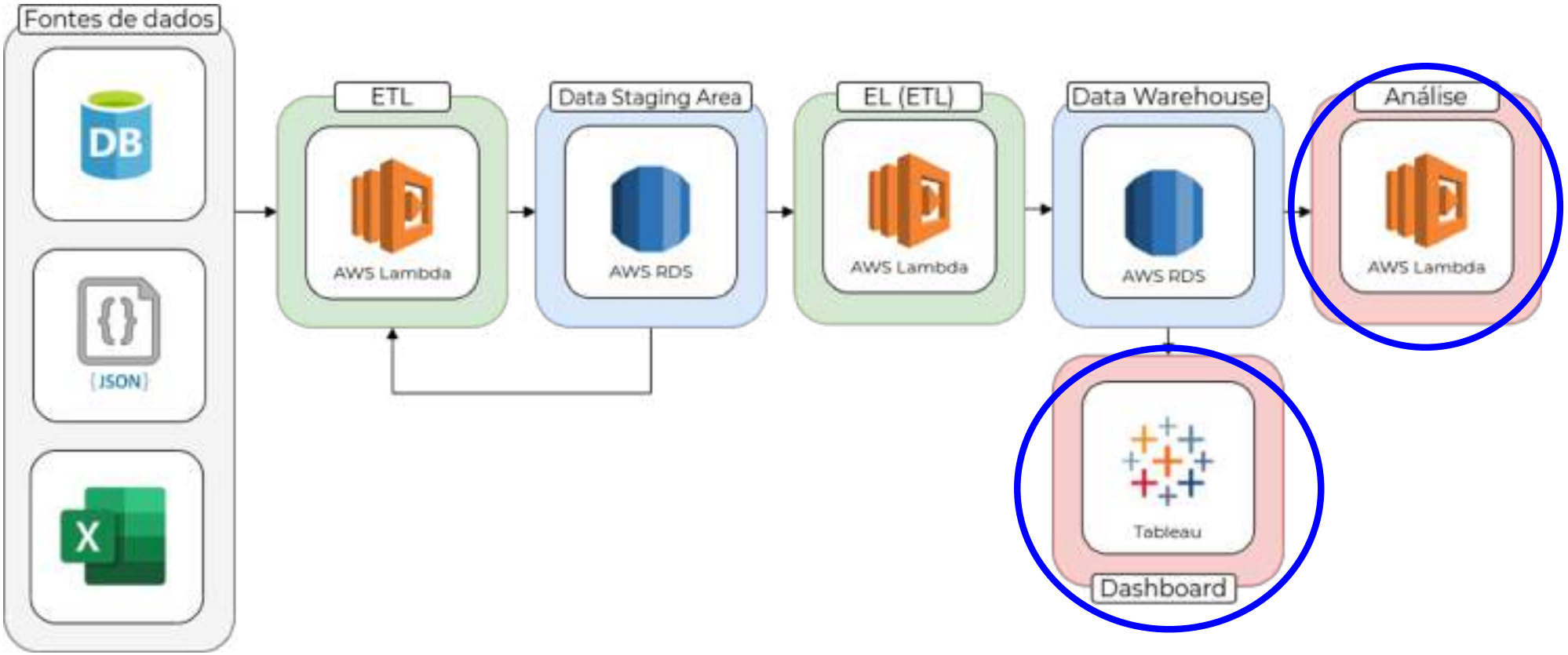
Processamento de Dados em Lote (Nuvem)



Data Staging Area e Data Warehouse

Serviço de provisionamento de bancos de dados relacionais em nuvem da AWS, chamado de AWS RDS (*Relational Database Service*)

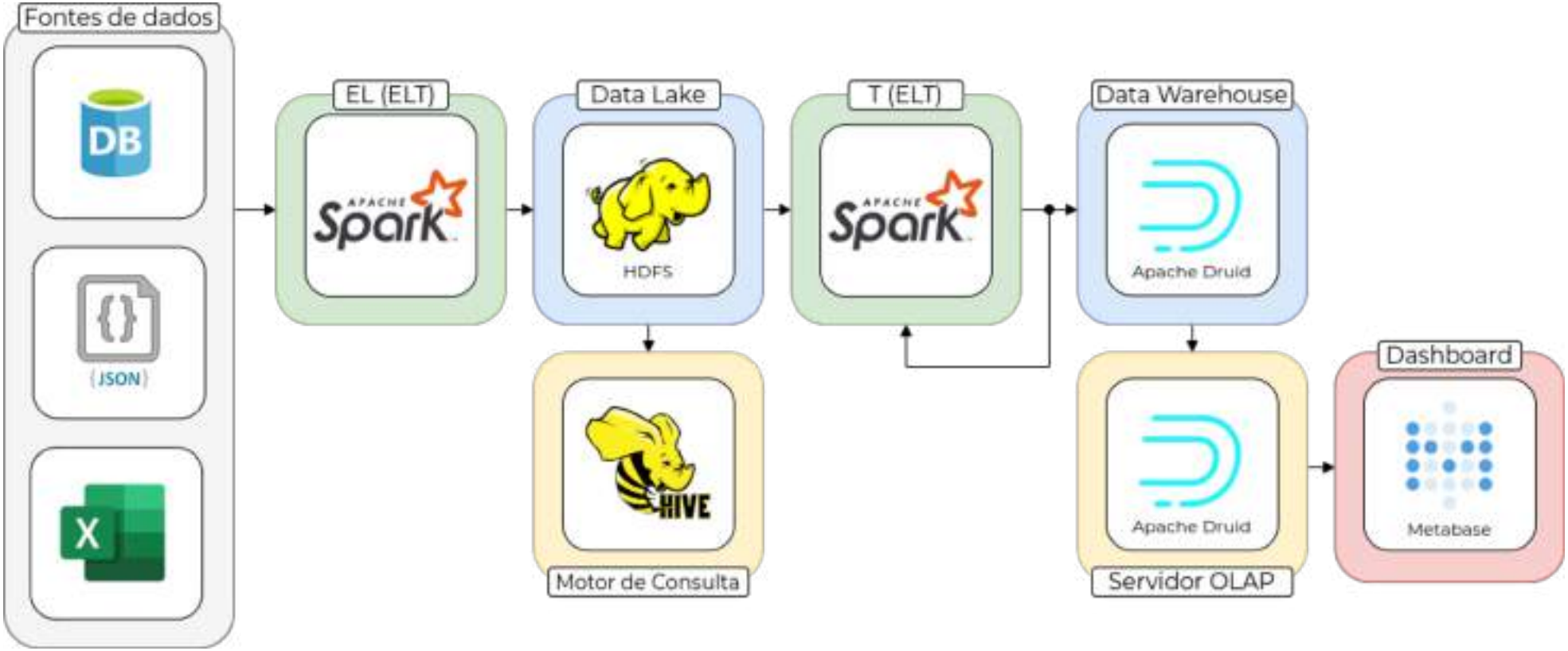
Processamento de Dados em Lote (Nuvem)



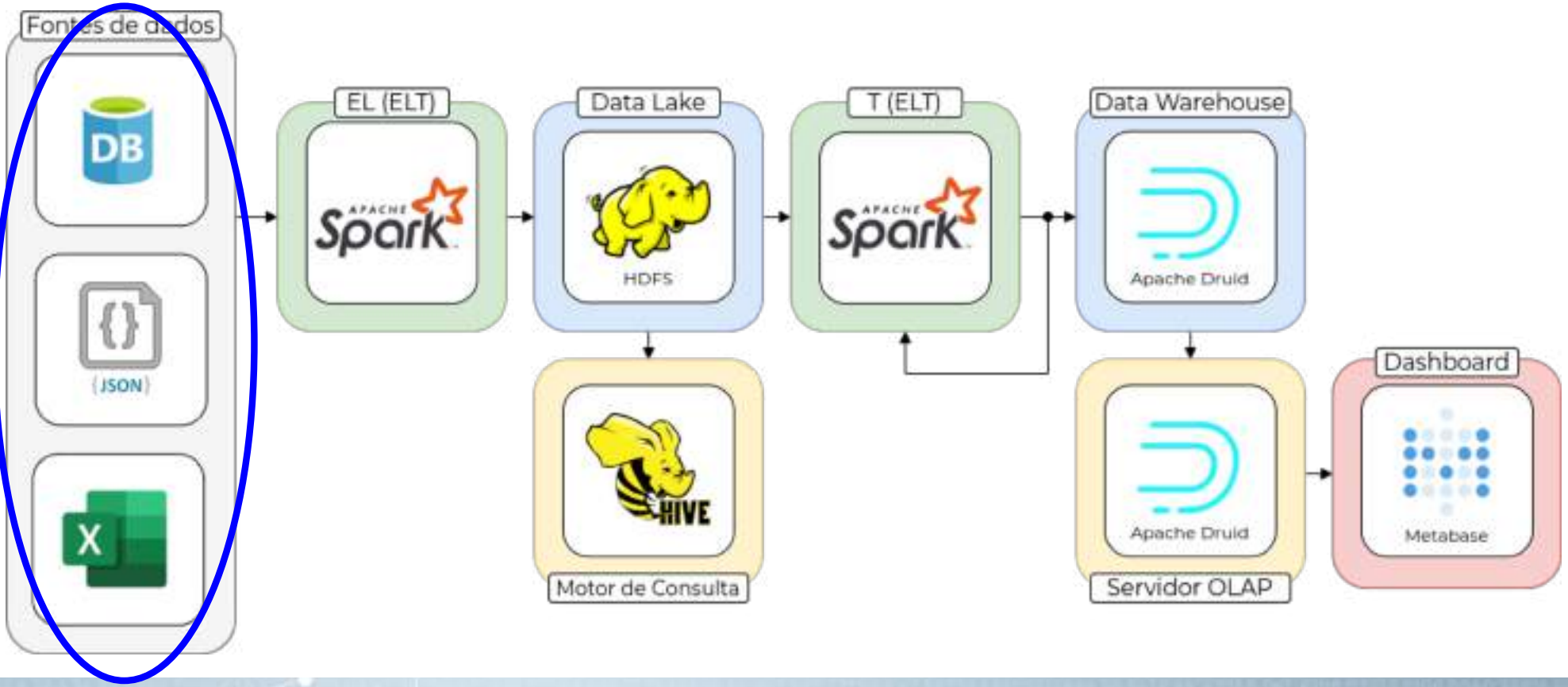
Exemplos de Pipeline

- Volumes de Dados Tradicionais
- Big Data
- Data Streaming

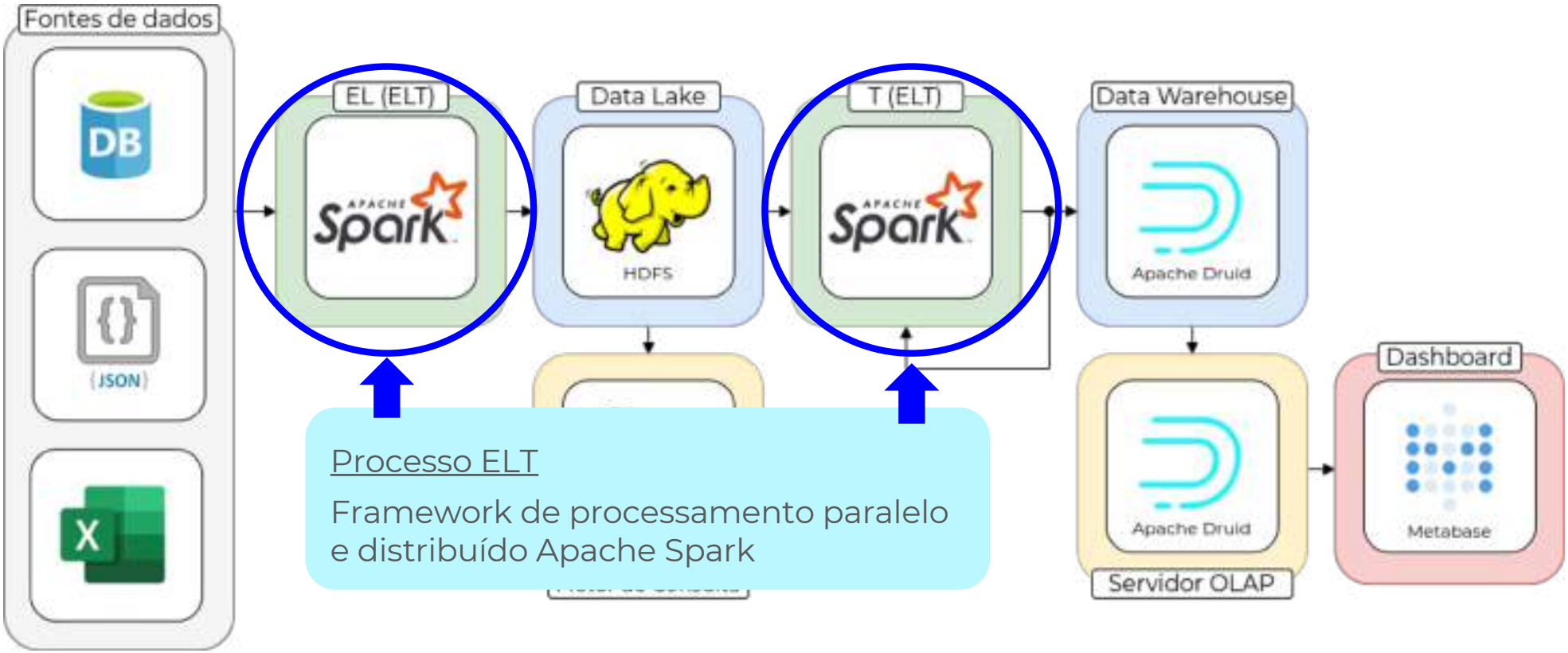
Processamento de Big Data em Lote



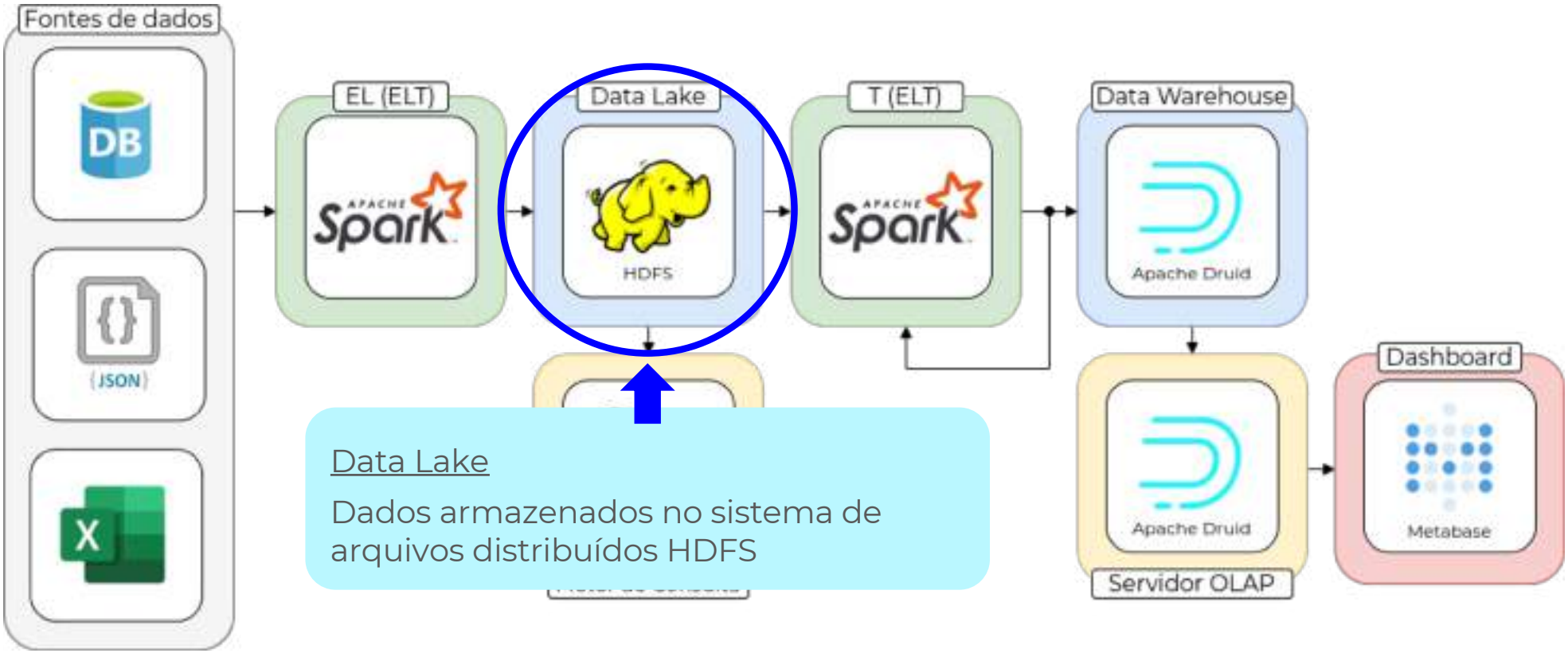
Processamento de Big Data em Lote



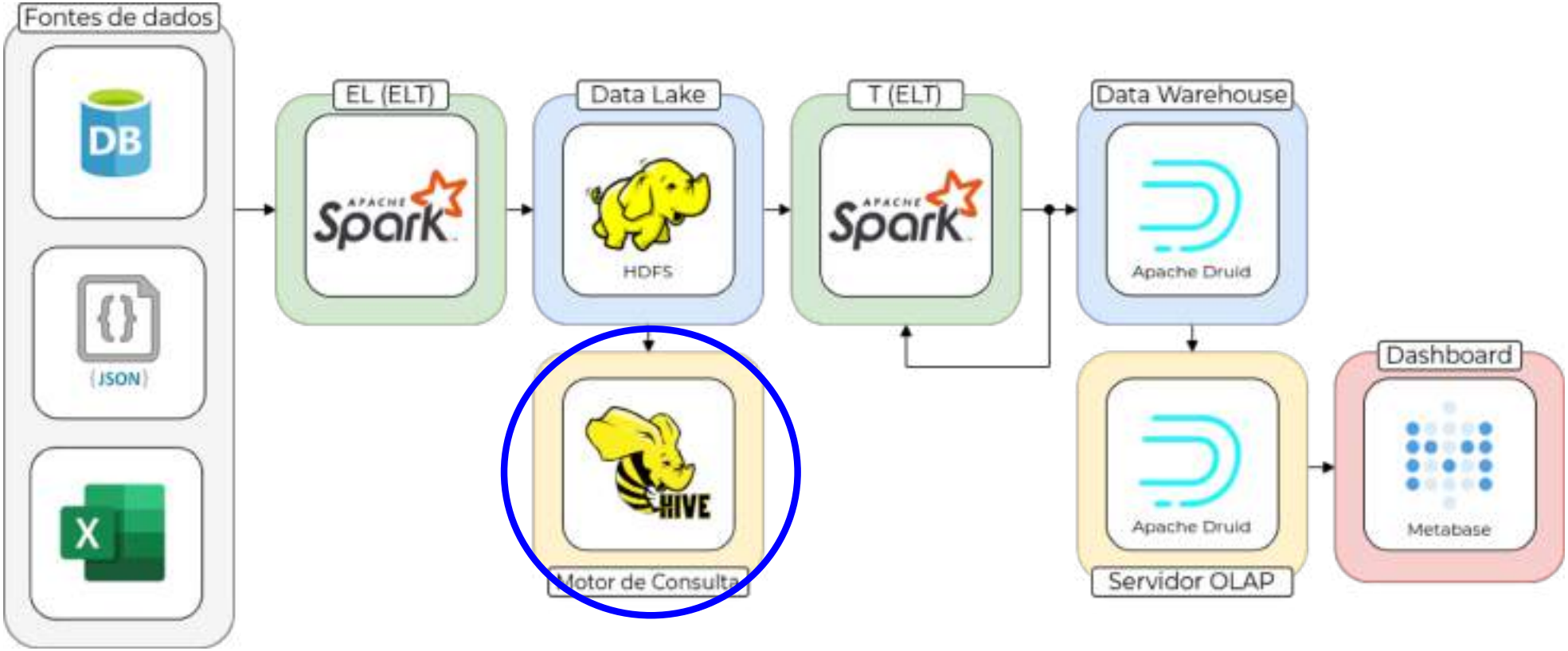
Processamento de Big Data em Lote



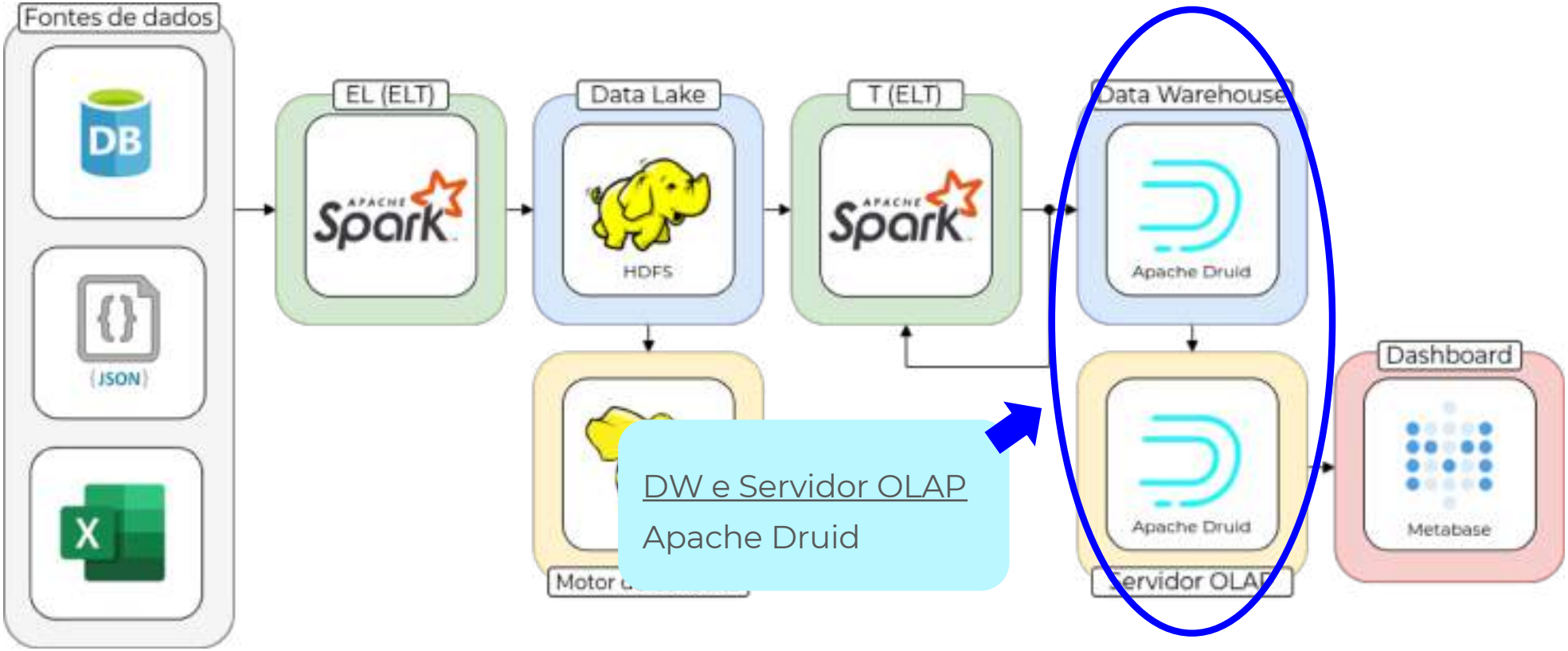
Processamento de Big Data em Lote



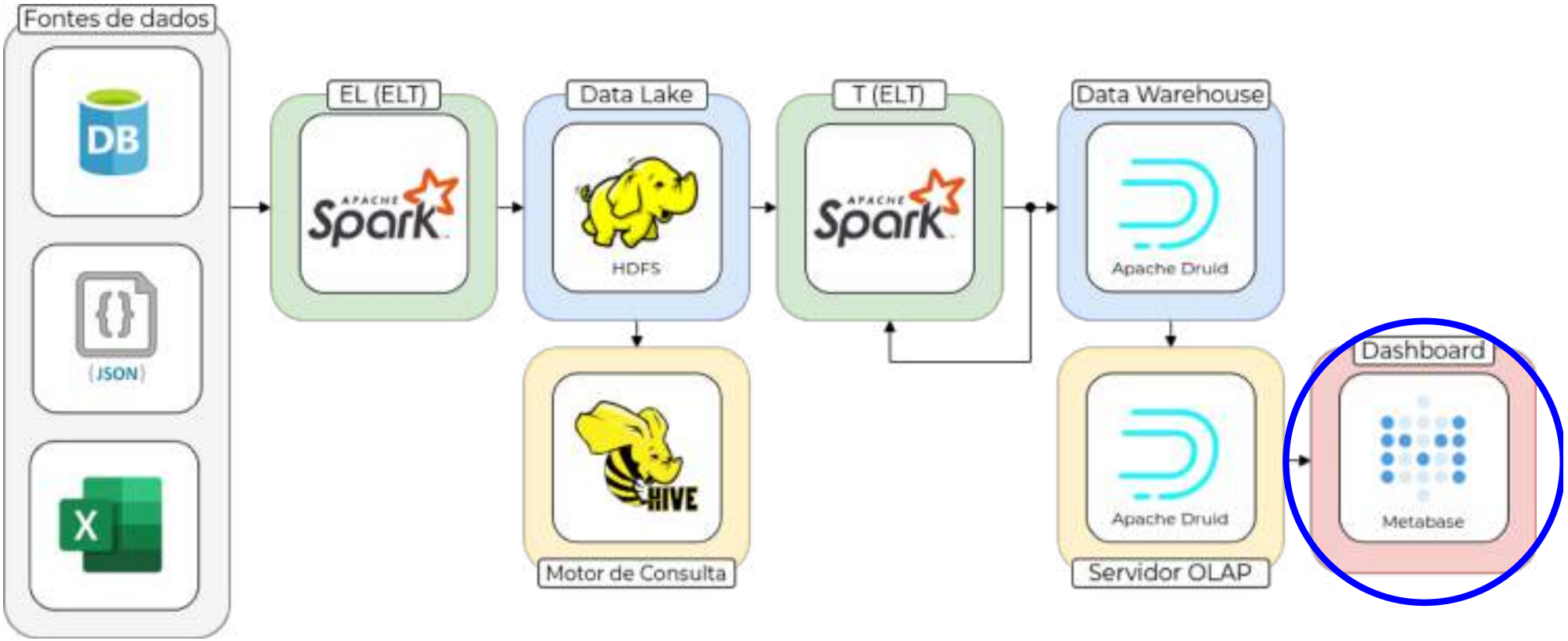
Processamento de Big Data em Lote



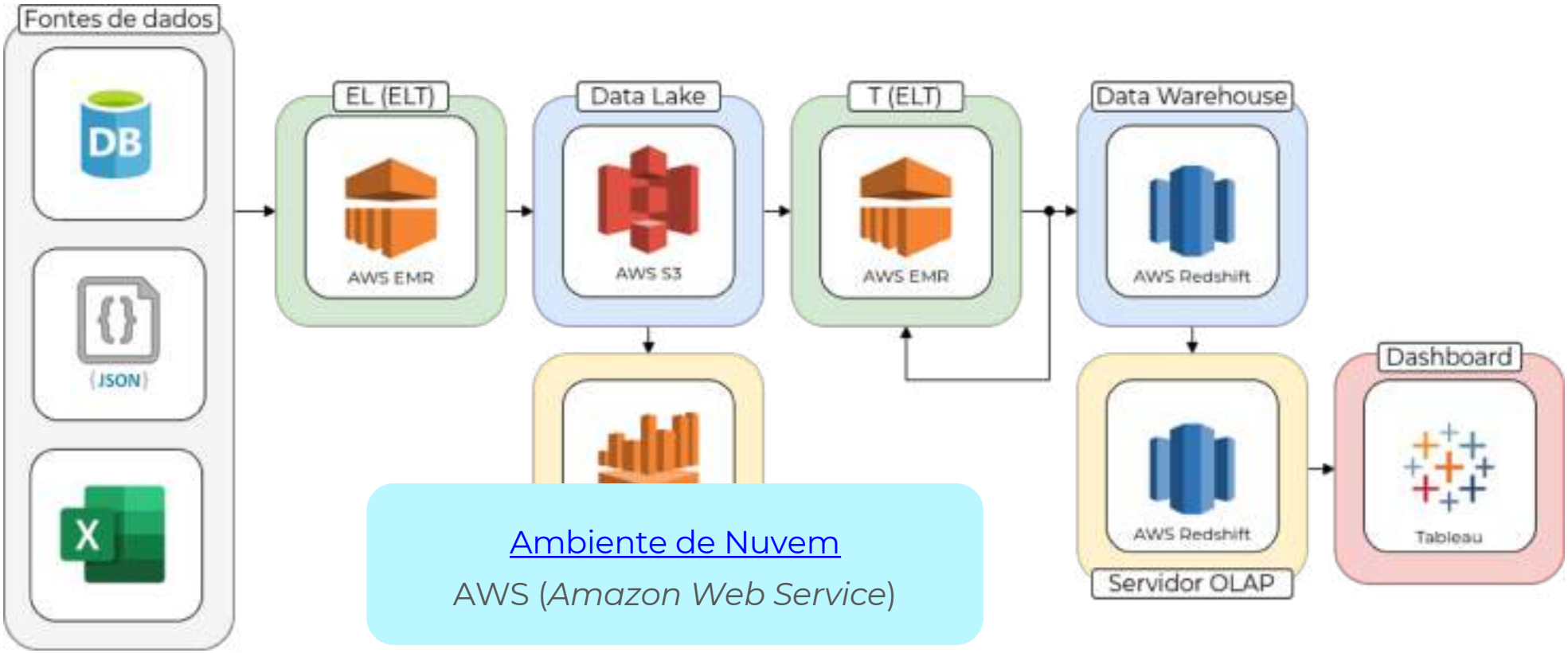
Processamento de Big Data em Lote



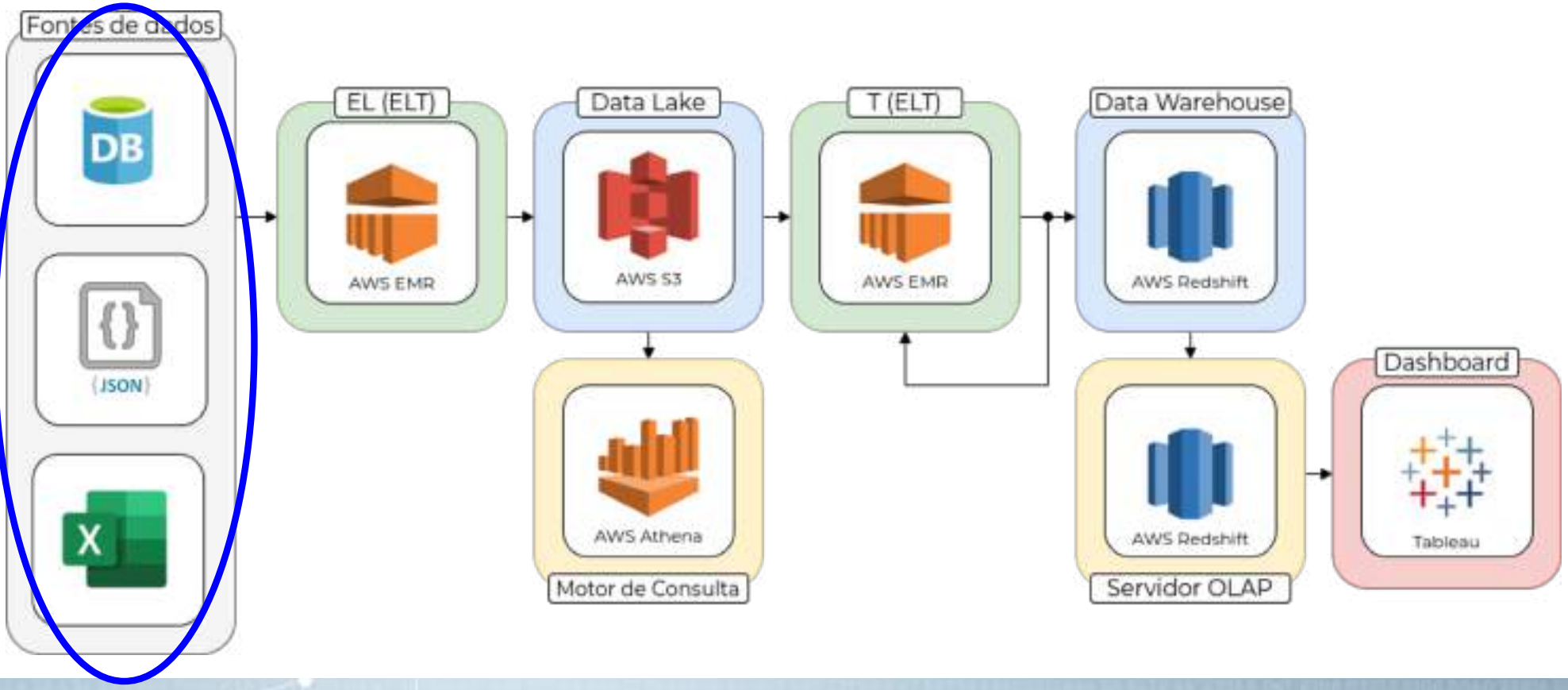
Processamento de Big Data em Lote



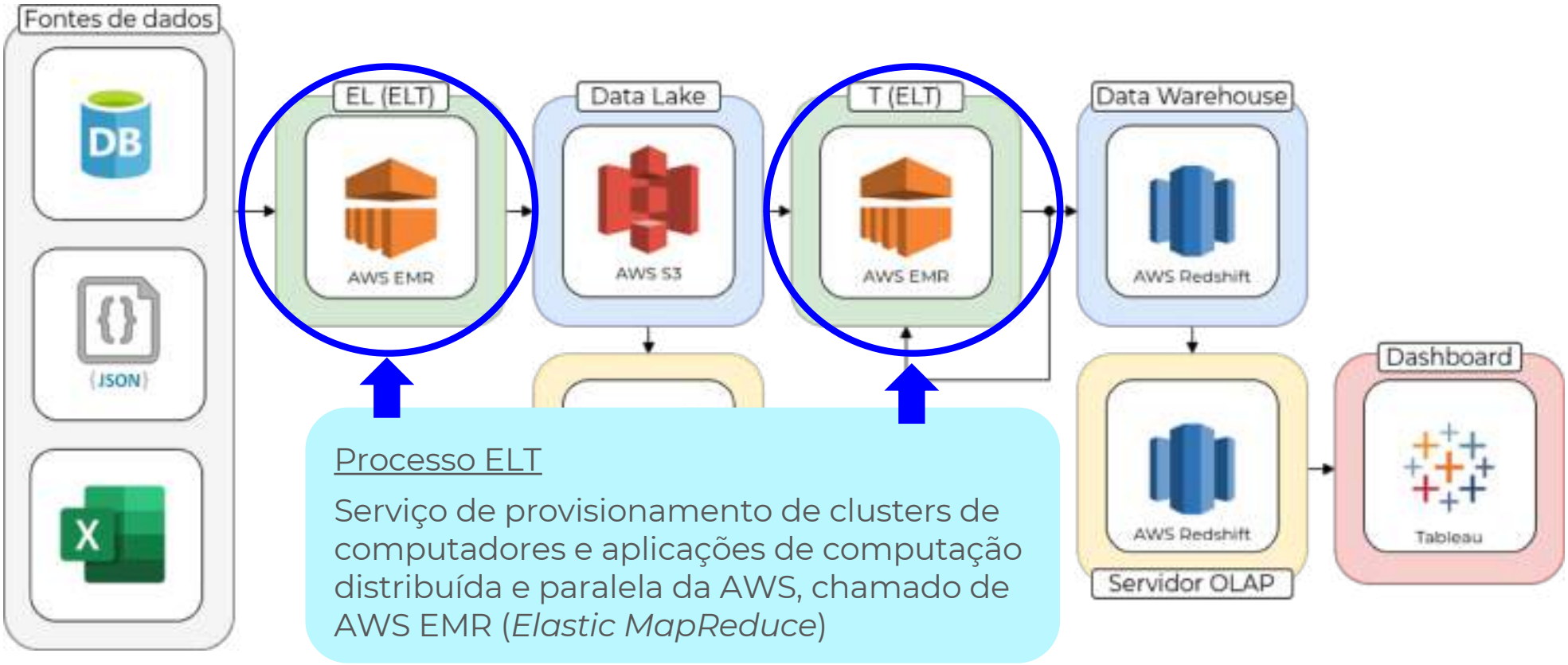
Processamento de Big Data em Lote (Nuvem)



Processamento de Big Data em Lote (Nuvem)



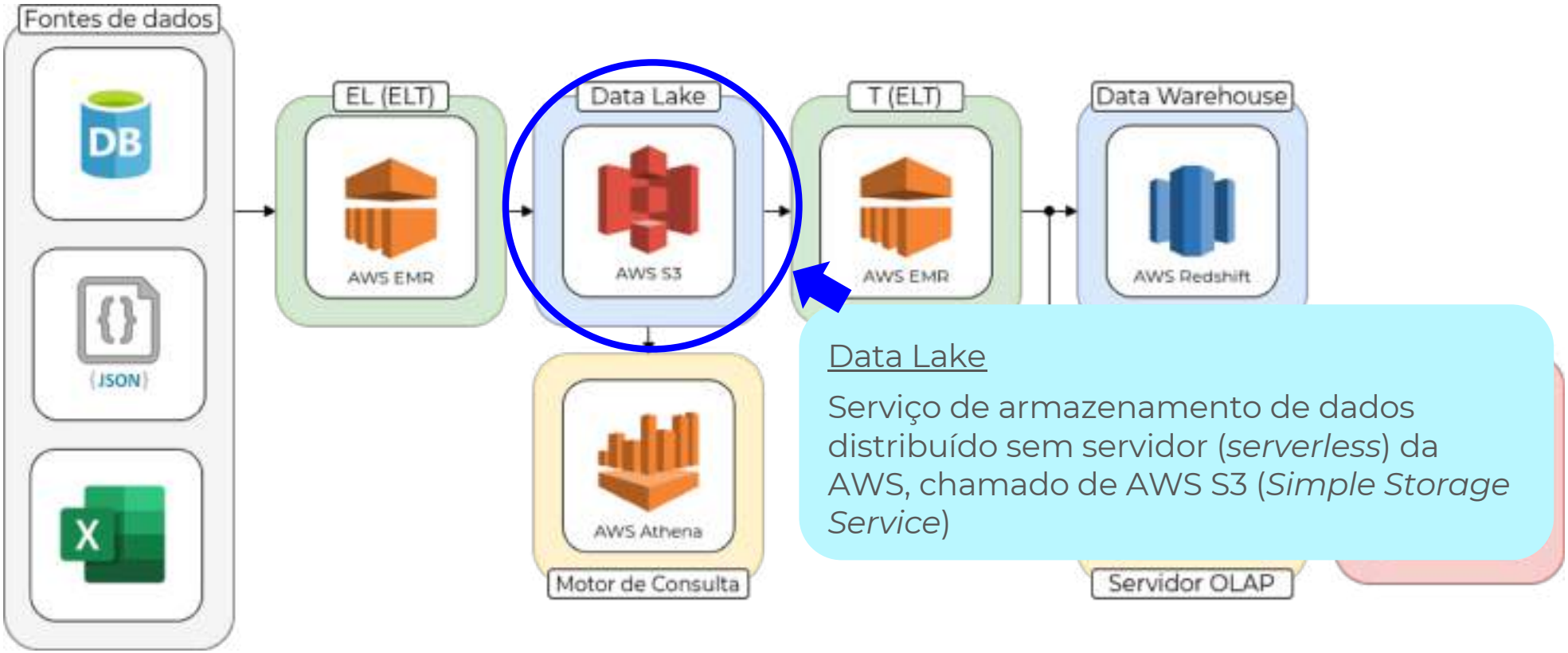
Processamento de Big Data em Lote (Nuvem)



Processo ELT

Serviço de provisionamento de clusters de computadores e aplicações de computação distribuída e paralela da AWS, chamado de AWS EMR (*Elastic MapReduce*)

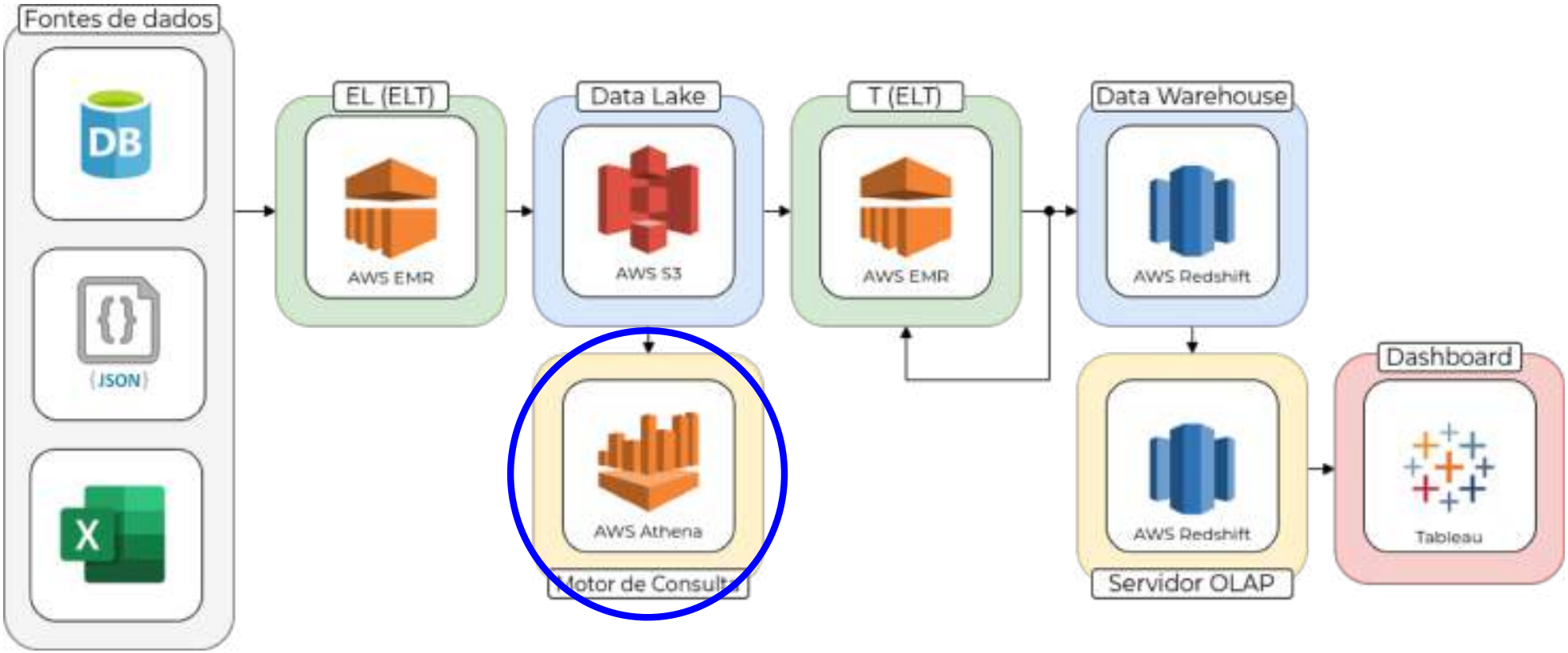
Processamento de Big Data em Lote (Nuvem)



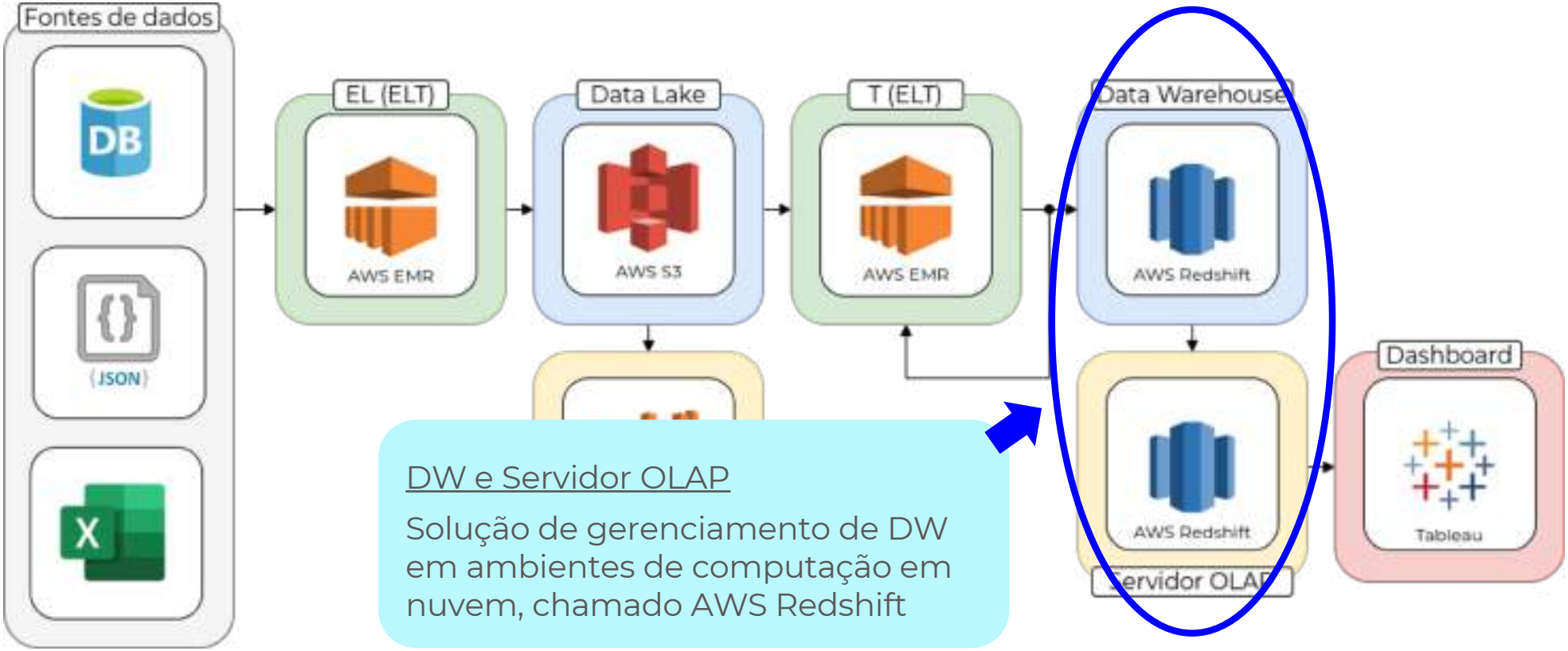
Data Lake

Serviço de armazenamento de dados distribuído sem servidor (*serverless*) da AWS, chamado de AWS S3 (*Simple Storage Service*)

Processamento de Big Data em Lote (Nuvem)



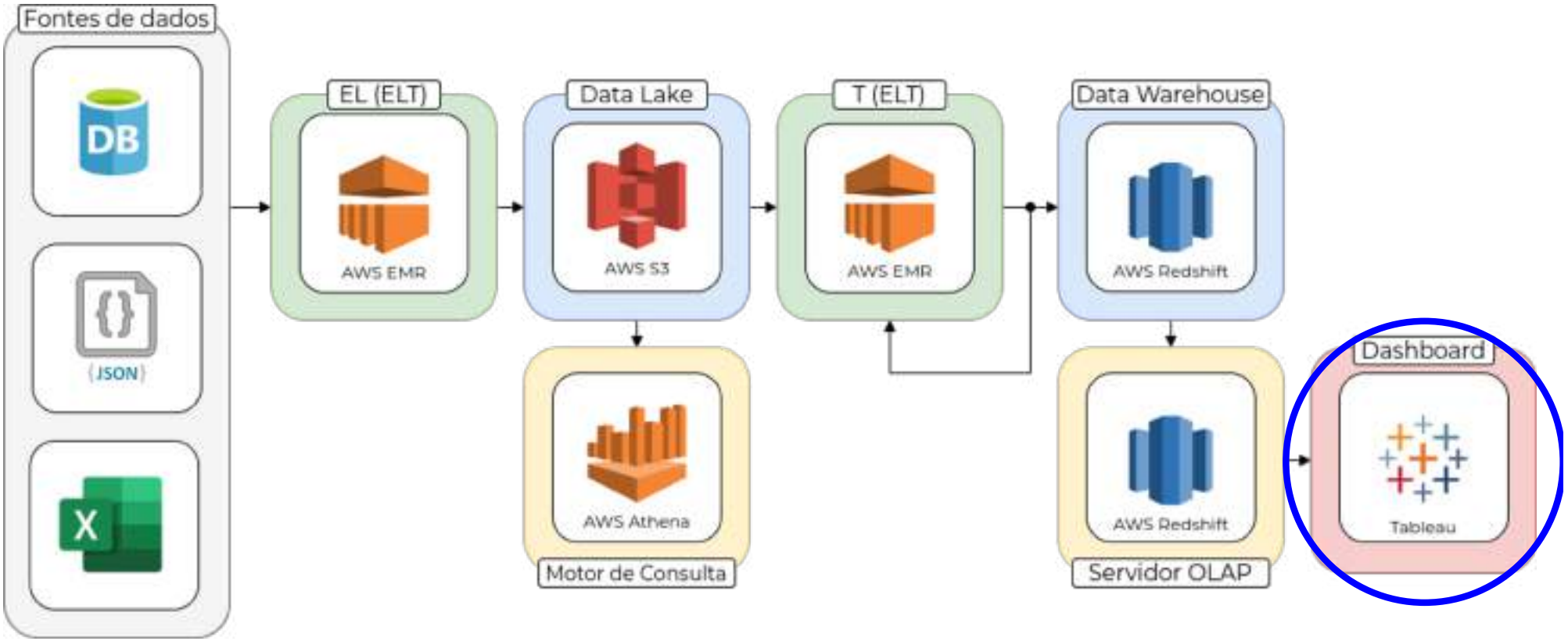
Processamento de Big Data em Lote (Nuvem)



DW e Servidor OLAP

Solução de gerenciamento de DW em ambientes de computação em nuvem, chamado AWS Redshift

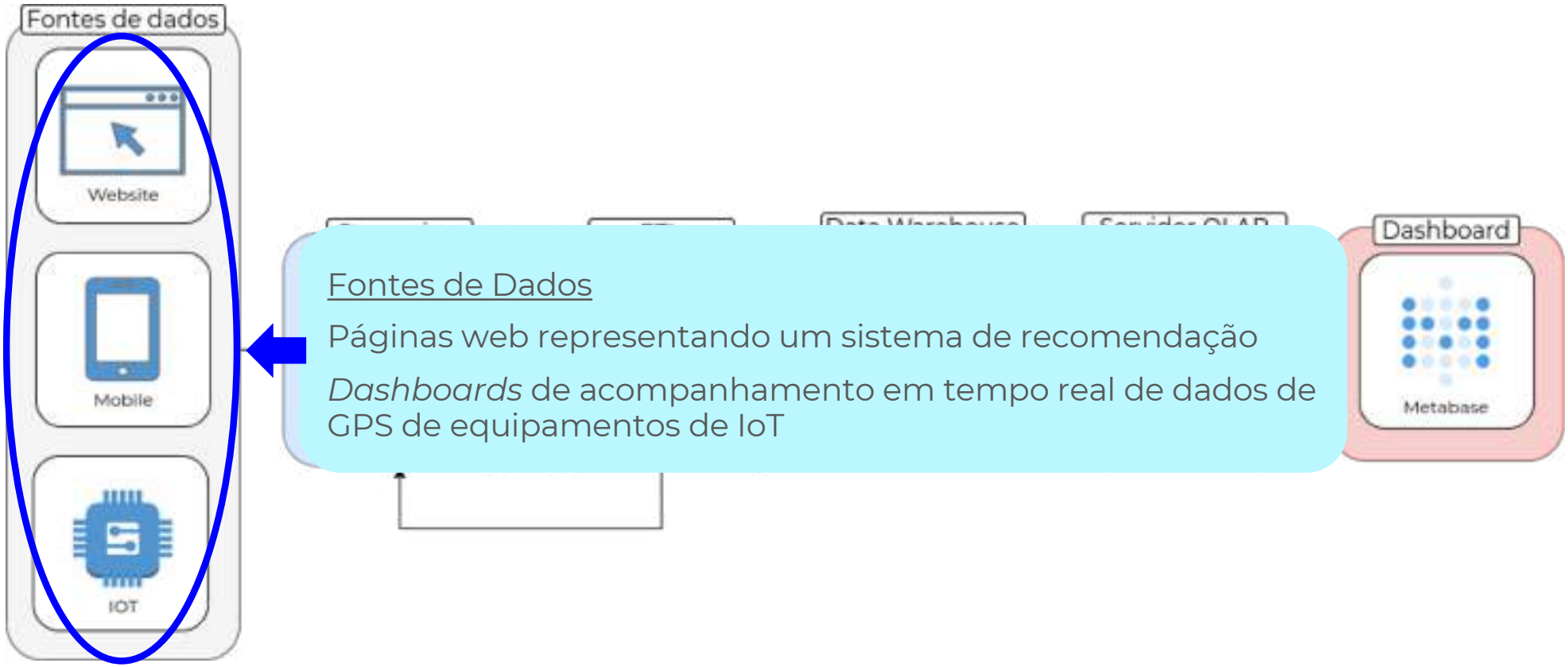
Processamento de Big Data em Lote (Nuvem)



Exemplos de Pipeline

- Volumes de Dados Tradicionais
- Big Data
- Data Streaming

Processamento de Streaming de Big Data



Fontes de Dados

Páginas web representando um sistema de recomendação

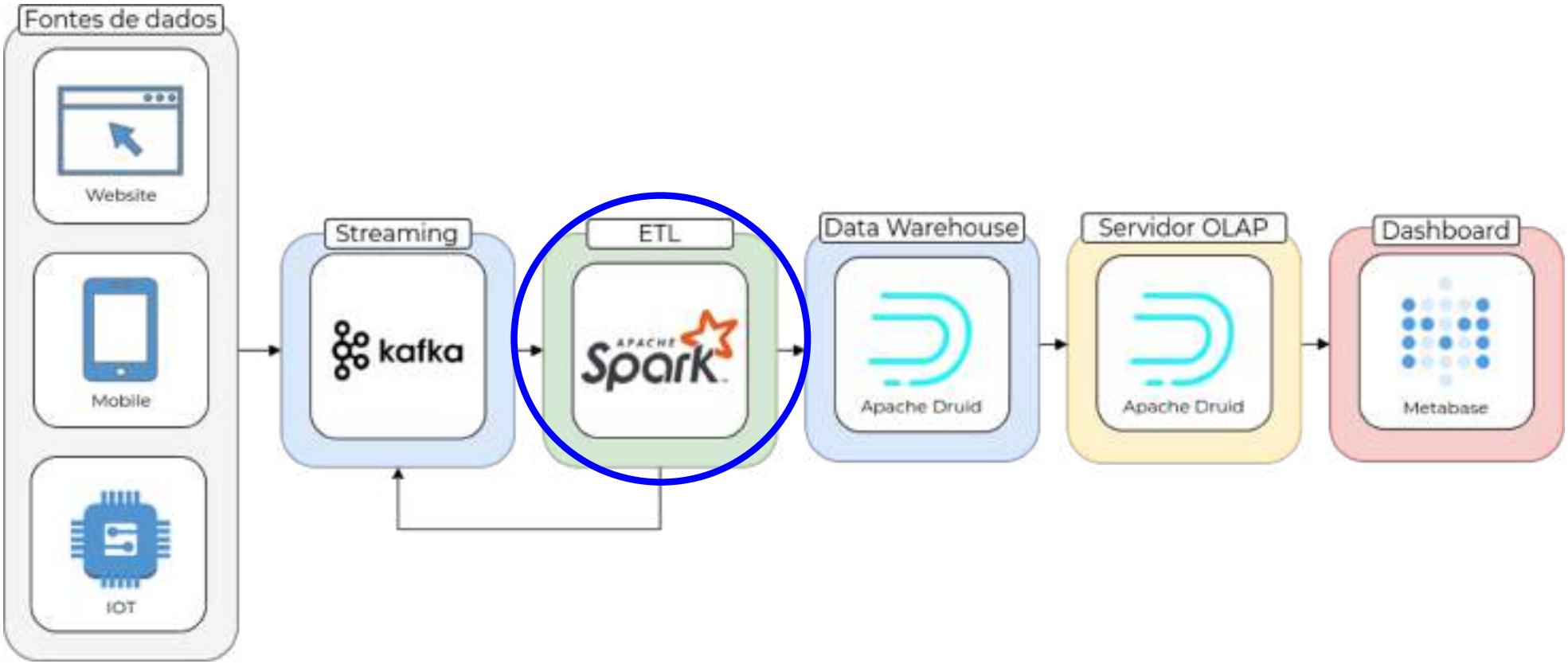
Dashboards de acompanhamento em tempo real de dados de GPS de equipamentos de IoT

Processamento de Streaming de Big Data

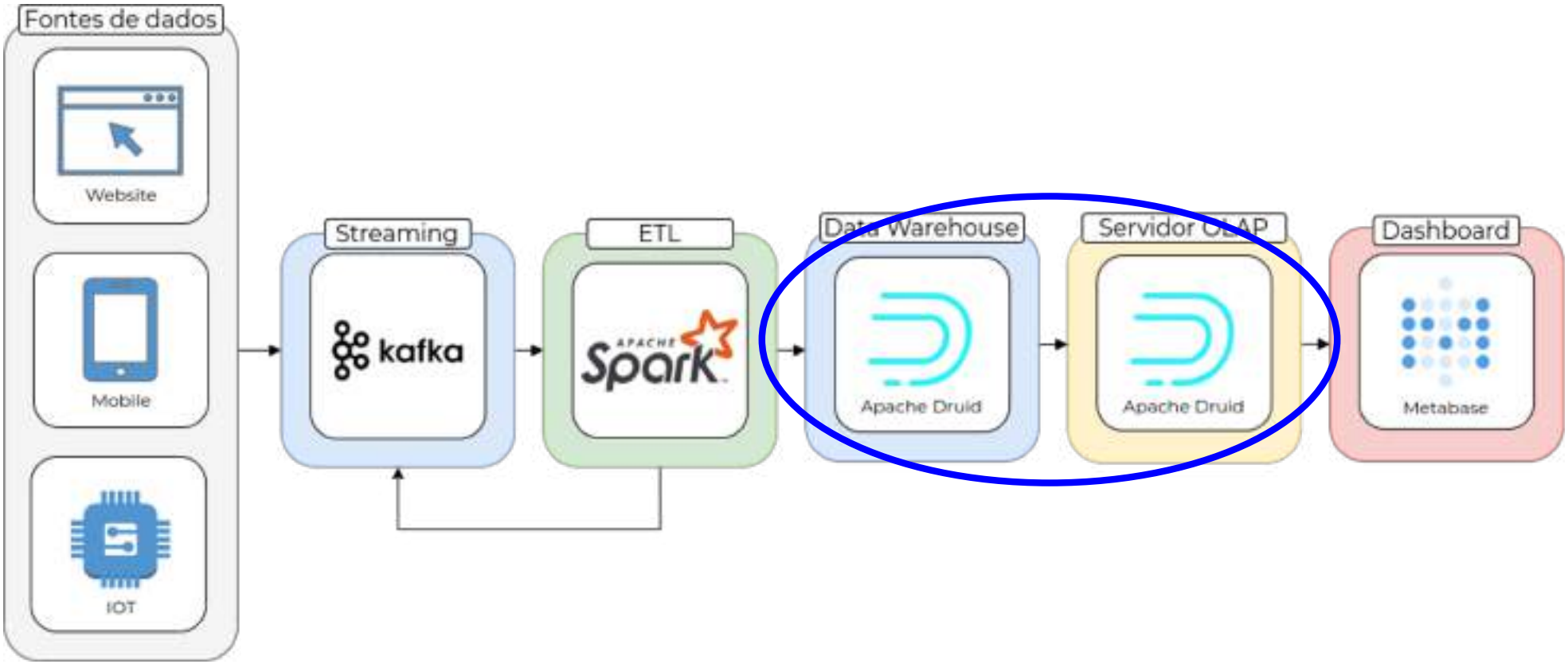
Streaming

Dados são obtidos assim que eles são gerados usando-se Apache Kafka

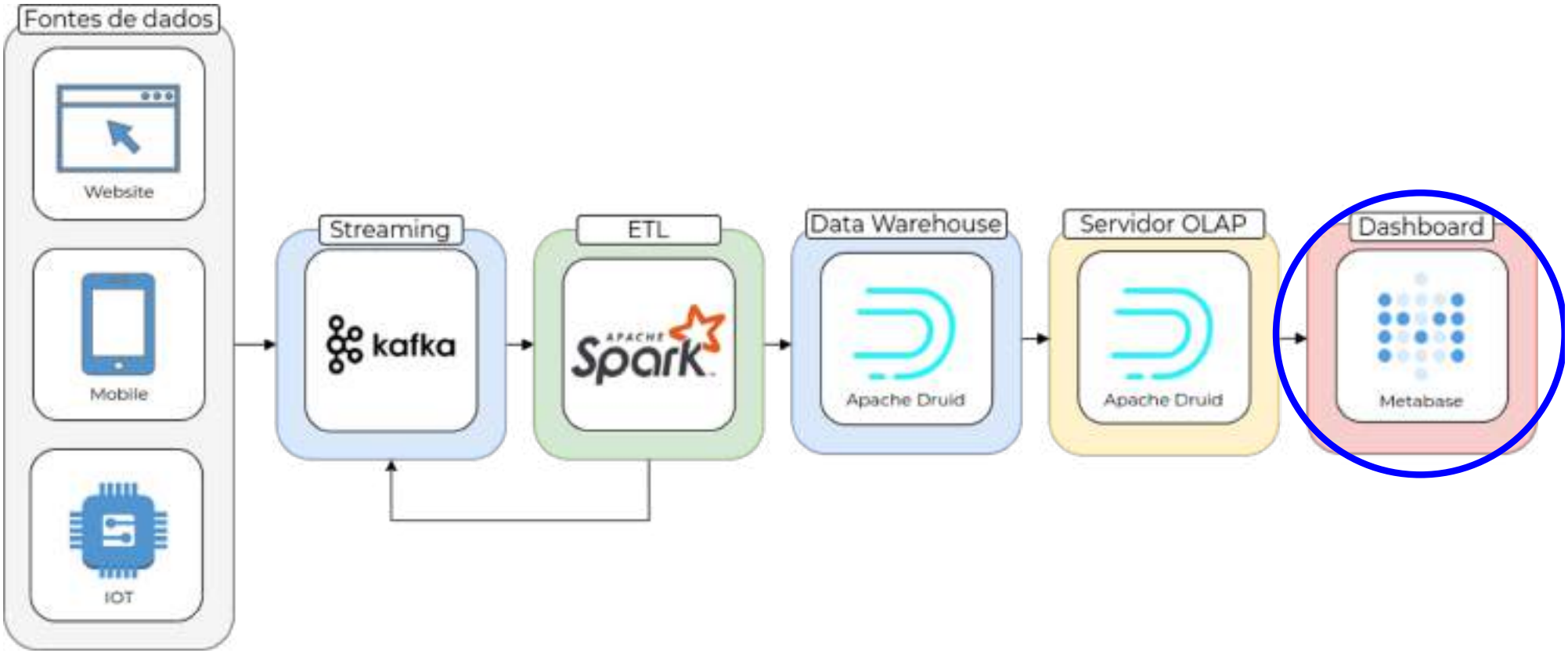
Processamento de Streaming de Big Data



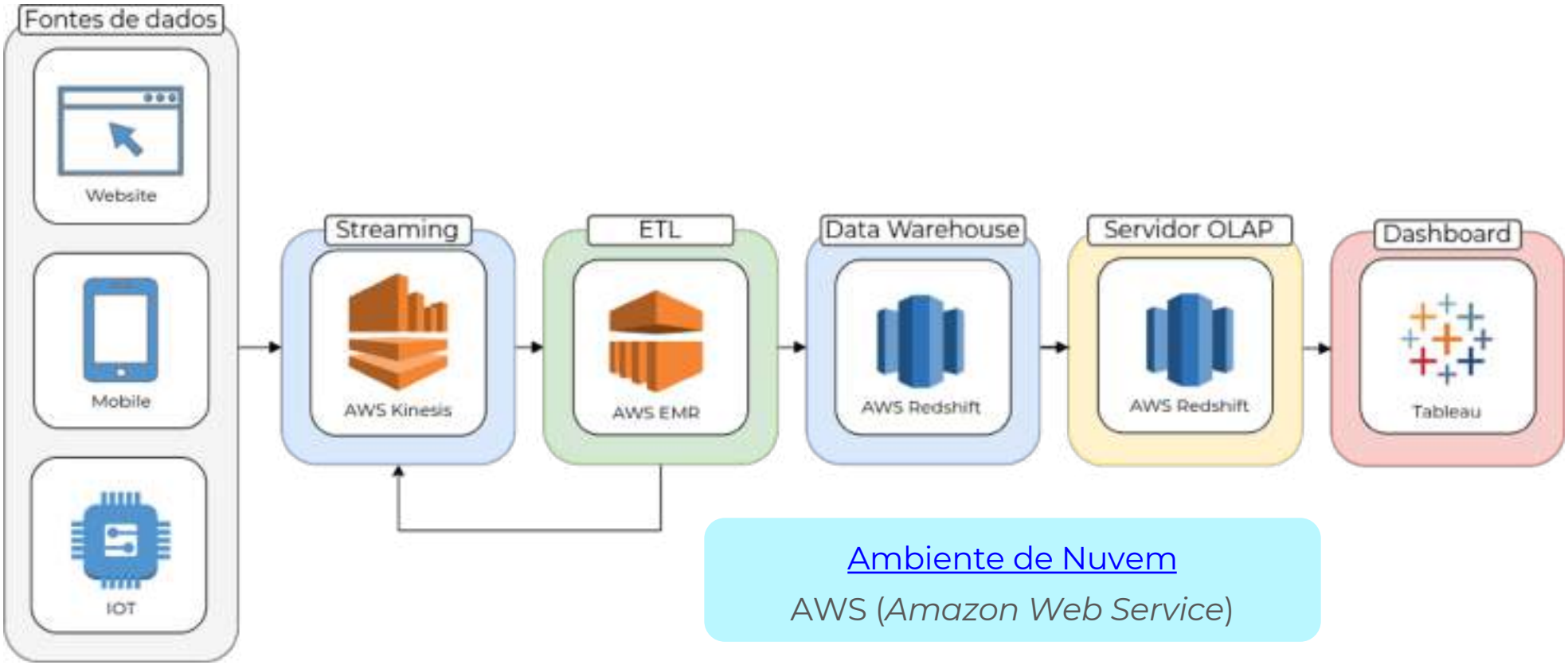
Processamento de Streaming de Big Data



Processamento de Streaming de Big Data



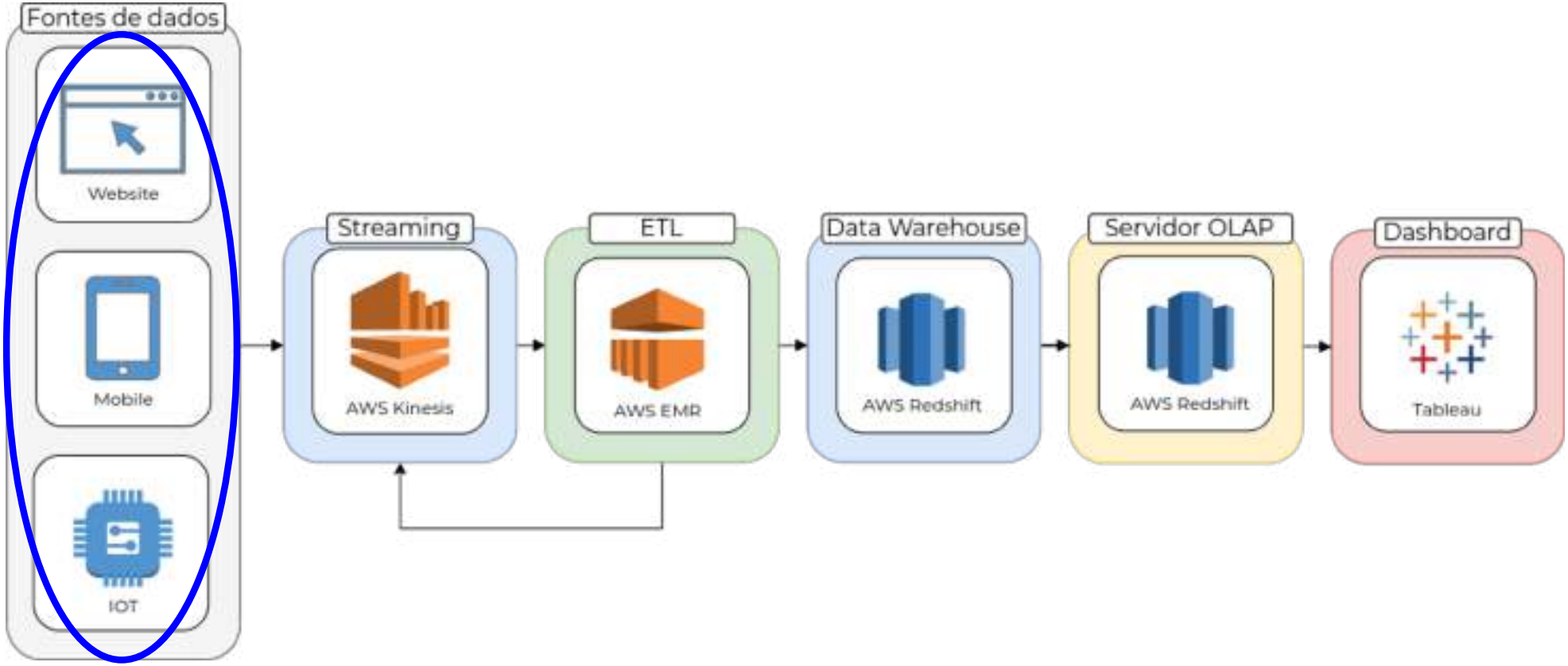
Processamento de Streaming de Big Data (Nuvem)



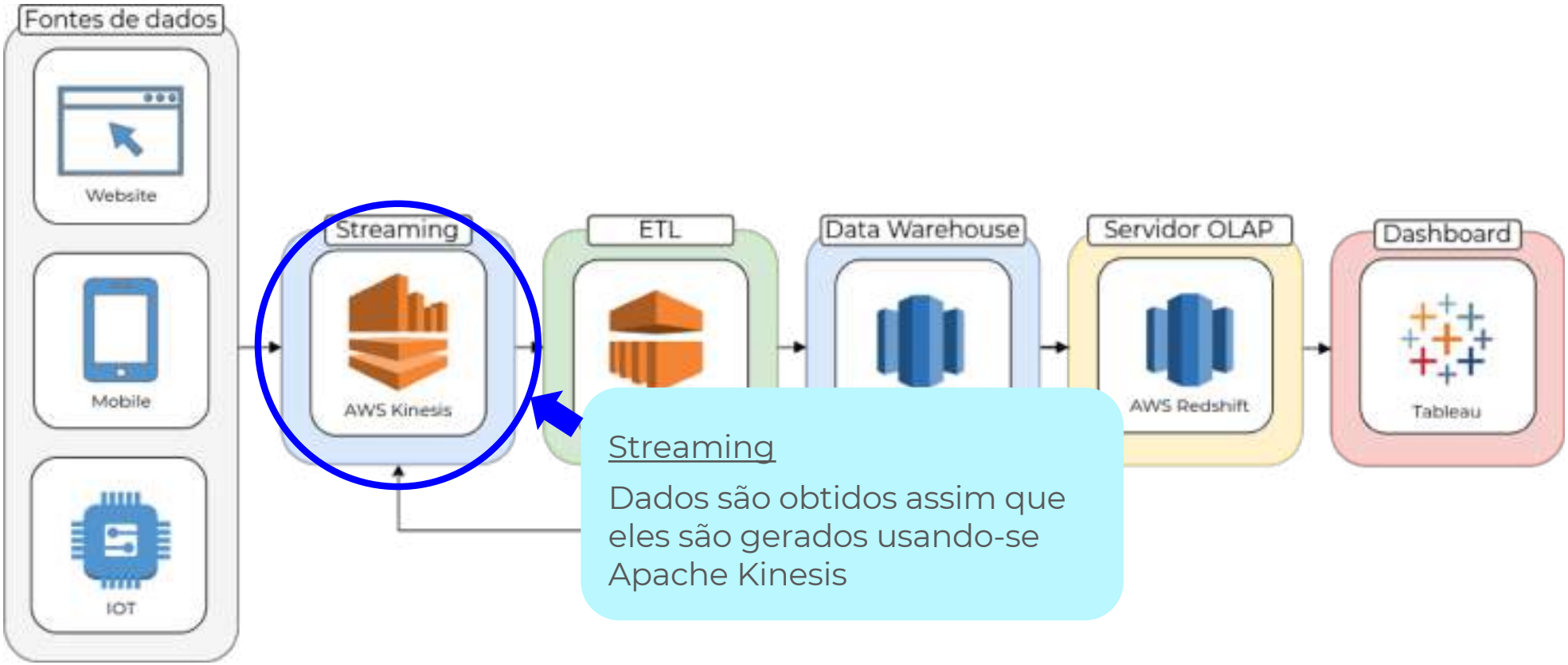
Ambiente de Nuvem

AWS (*Amazon Web Service*)

Processamento de Streaming de Big Data (Nuvem)



Processamento de Streaming de Big Data (Nuvem)

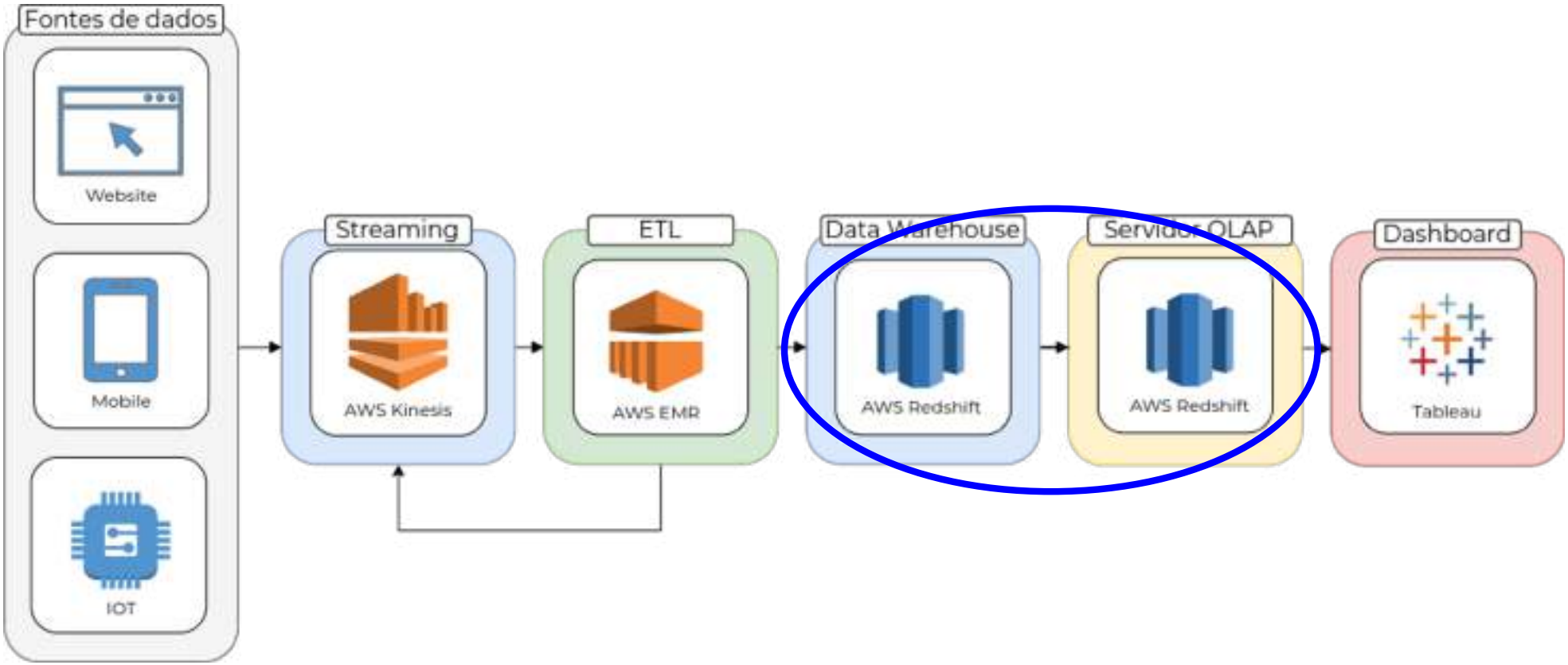


Streaming

Dados são obtidos assim que eles são gerados usando-se Apache Kinesis

Processamento de Streaming de Big Data (Nuvem)

Processamento de Streaming de Big Data (Nuvem)



Processamento de Streaming de Big Data (Nuvem)