Curso 3: Administração de Dados Complexos em Larga Escala

-- Mais conceitos de DW --

Prof. Jose Fernando Rodrigues Junior

Objetivo: apresentar os conceitos de ETL e OLAP

Arquitetura de um data warehouse

- Arquitetura definida pelo contexto da organização

- De maneira geral, tem as seguintes camadas:
  1. ***Operacional (*OLTPs*)*:** fornecem dados
  2. ***De acesso aos dados*:** ETL
  3. ***Acesso à informação*:** ferramentas de acesso a dados, geração de relatórios, e análise (**OLAP**) ☐ **Business Intelligence**
  4. ***Metadados***: detalhamento do conteúdo do data warehouse ☐ **dicionário de dados**

ETL – Extract Transform Load

- 1) Recolher (**extrair**) os dados não importando qual o tipo do sistema de dados;
- 2) Padronizar (**transformar**) os dados, para terem um significado comum mesmo que, originalmente, codificados de maneira diferente; resolução de dados ausentes e espúrios;
- 3) Unir (**carregar**) os resultados das duas operações em um único sistema capaz de responder às minhas perguntas.

ETL – Extract Transform Load

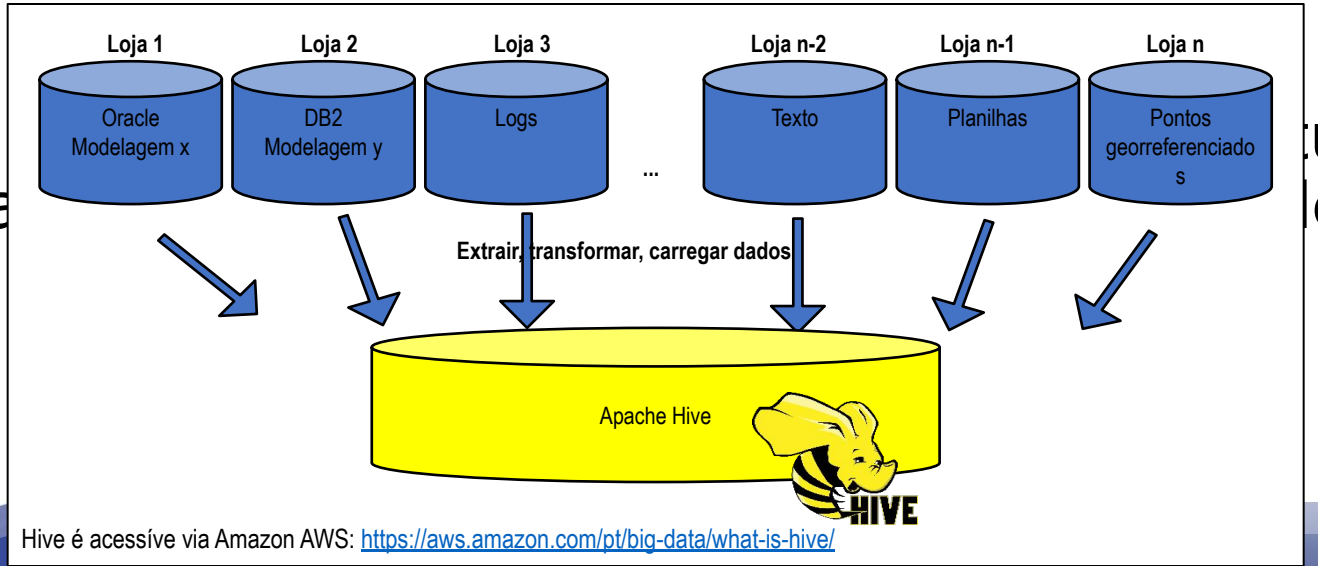
- O processo de se extrair (**Extract**), transformar (**Transform**) e carregar (**Load**) os dados a partir das diversas fontes de dados é denominado ETL

- ETL – uma das camadas principais da arquitetura de um Data Warehouse ☐ **Consolidação de dados**

ETL – Extract Transform Load

- O processo de se extrair (**Extract**), transformar (**Transform**) e carregar (**Load**) os dados a partir das diversas fontes de dados é denominado ETL

- ETL – ...tura de um Da... **...os**



Hive é acessíve via Amazon AWS: https://aws.amazon.com/pt/big-data/what-is-hive/

ETL – Extract Transform Load

- **Extração/transformação** (**Extract/Transform**) de dados
  - extração de **múltiplas fontes**
  - **consolidação e integração** de dados de múltiplas fontes
  - **limpeza e validação**
  - **conversão dos dados** para o modelo do DW

ETL – Extract Transform Load

Transformação (**Extract/Transform**)

- **Seleção** de apenas determinadas colunas para carregar
- **Codificação** de valores categóricos (ex.: gênero)
- Alteração do **separador** de valores (ex.: tsv ☐ csv)
- **Derivação** de um novo valor calculado (montante_vendas = qtde * preço_unitário, por exemplo)
- **Junção** de dados provenientes de diversas fontes
- Geração de valores de **chaves substitutas** (surrogate keys - id)
- **Transposição ou rotação**, transformando múltiplas colunas em múltiplas linhas ou vice-versa
- **Quebra de uma coluna** em diversas colunas

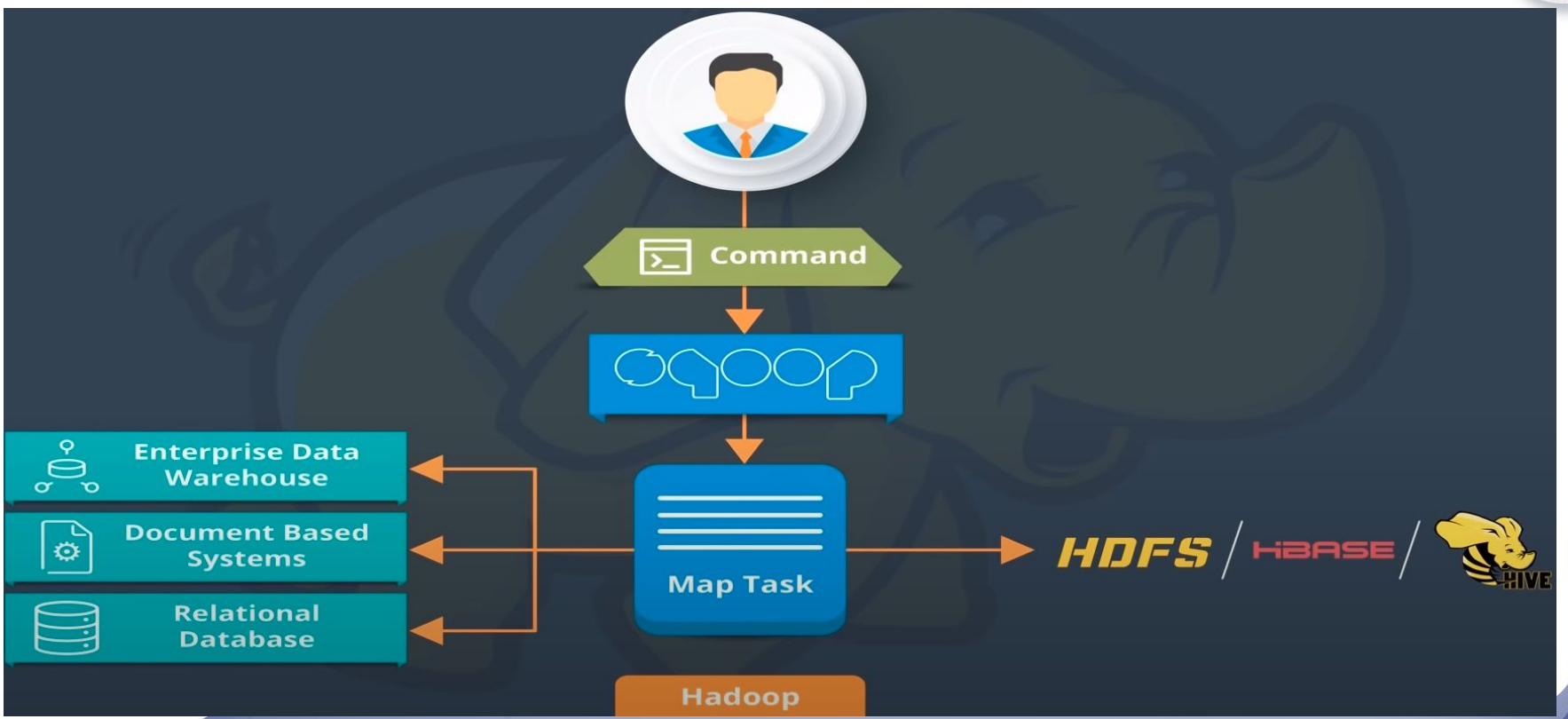
ETL – Extract Transform Load

- Carregamento (**Load**) de dados
  - armazenamento de acordo com o **modelo do DW**
  - criação e manutenção de **estruturas de dados**
  - criação e manutenção de **caminhos de acesso**
  - **tratamento de dados** que variam no tempo
  - **suporte a atualização**
    - *refresh*
    - *purging* (eliminação)

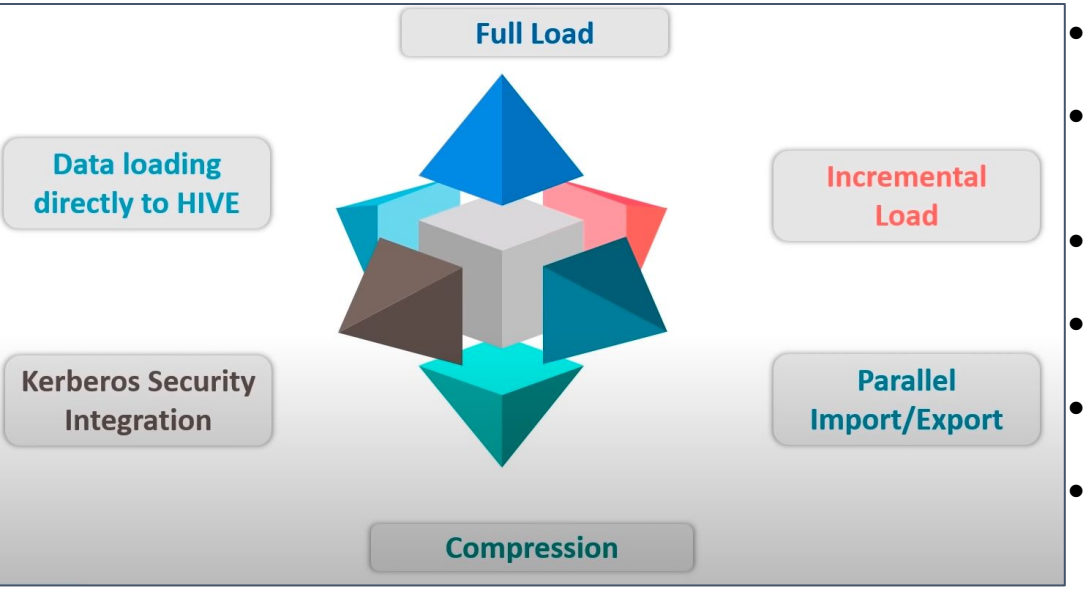
Load via Apache Sqoop

- No ecossistema Hadoop, é possível fazer o loading a partir de bases de dados relacionais por meio da ferramenta **Apache Sqoop**;
- O Apache Sqoop suporta a **transferência bidirecional** de dados entre o Hive e os principais SGBDs: MySQL, Informix, PostgreSQL, Oracle, IBM DB2 e Netezza, entre outros;
- Para tanto, basta **definir um schema dentro do Hive**, e usar os comandos de importação.

Load via Apache Sqoop

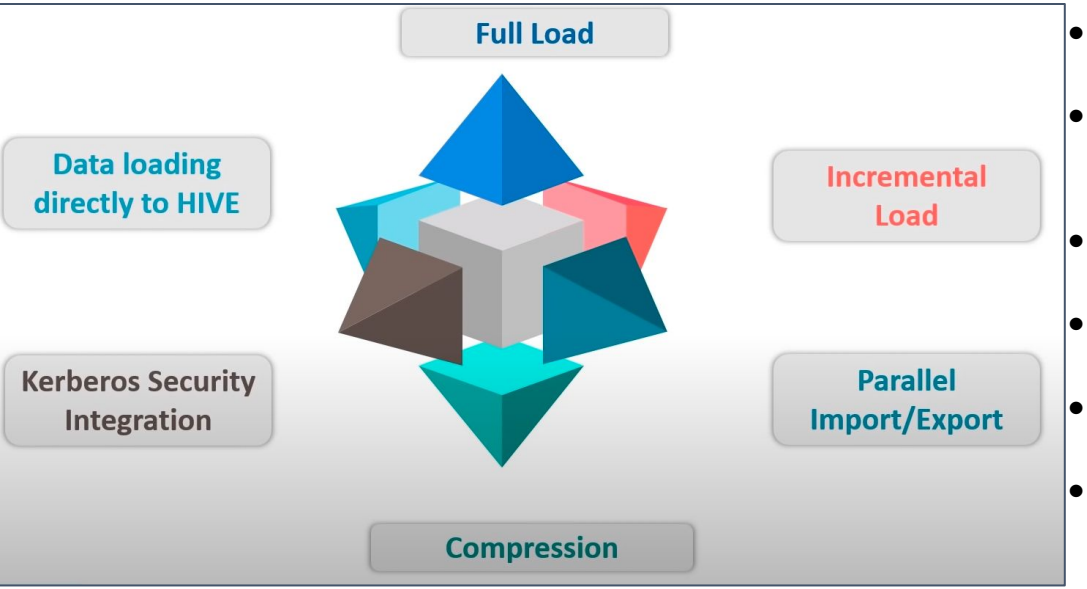


Load via Apache Sqoop



- **Full load**: tabelas inteiras carregadas com um único comando;
- **Incremental load**: carregamento apenas de dados que passaram por update;
- **Paralelismo**: beneficia-se do paralelismo do HDFS;
- **Compression**: dados comprimidos para importar e exportar;
- **Kerberos**: autenticação de segurança;
- **Carregamento direto** para Hive ou HBase.

Load via Apache Sqoop



- **Full load**: tabelas inteiras carregadas com um único comando;
- **Incremental load**: carregamento apenas de dados que passaram por update;
- **Paralelismo**: beneficia-se do paralelismo do HDFS;
- **Compression**: dados comprimidos para importar e exportar;
- **Kerberos**: autenticação de segurança;
- **Carregamento direto** para Hive ou HBase.

Exemplo: para carregar todas as tabelas de uma base de dados MySQL Futebol para uma base de dados Hive:

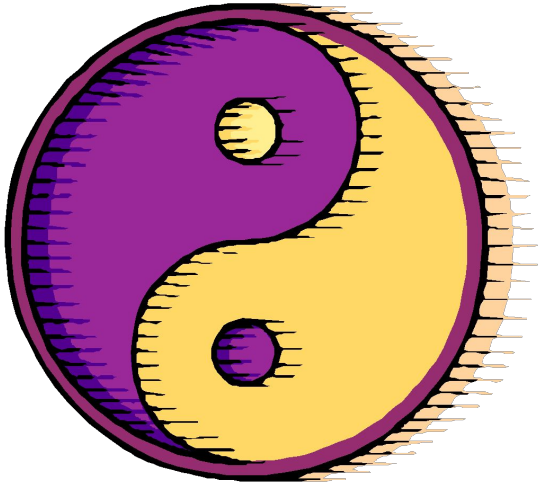
```
sqoop import-all-tables --conect jdbc:mysql/localhost/Futebol --username jose_rodrigues
```

Data warehouse vs Banco de dados operacional

Data warehouse	Banco de dados operacional
Orientado à **análise, estático**	Orientado a **transações, dinâmico**
Grande (centenas de GBs até TBs)	**Pequeno/Médio** (MBs até alguns GBs) – distribuído se necessário
Dados **históricos**	Dados **correntes**
De-normalizado (poucas tabelas com muitas colunas)	**Normalizado** (muitas tabelas com poucas colunas)
Atualizações em ***Batch***	Atualizações **contínuas**
Otimizado para **acesso**	Otimizado para **escrita/atualização**

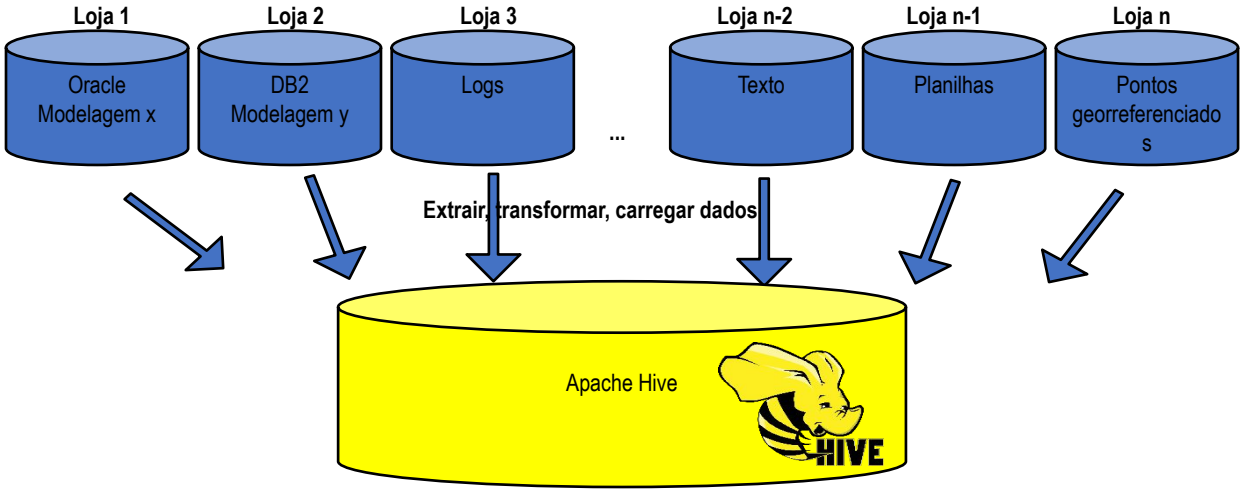
Juntos data warehouse e bancos de dados provém uma solução completa

Bancos de dados
Inserção/Atualização
(OLTP)



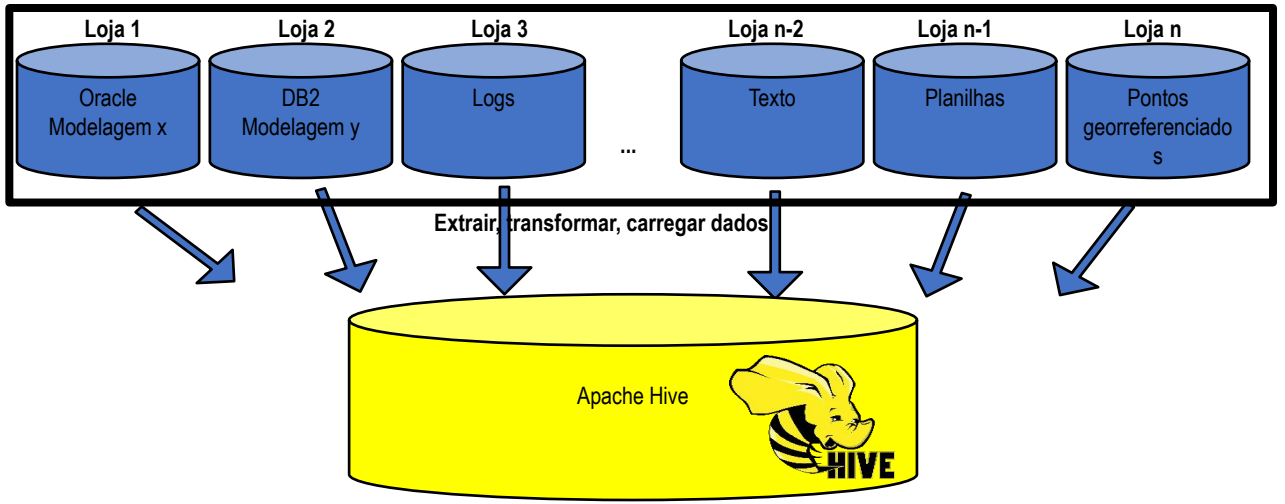
Data Warehouse
Acesso aos dados
(OLAP)

Visão Geral



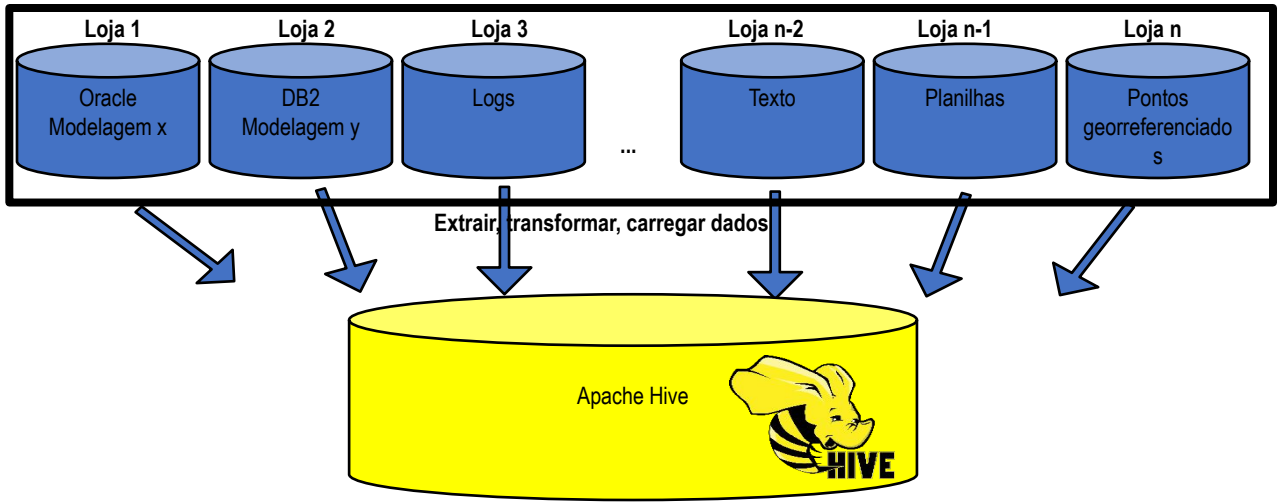
1. Camada Operacional (OLTPs)
2. Camada de acesso aos dados (ETL)
3. Camada de acesso à informação: Mineração de Dados, relatórios, OLAP
   ⇒ ***Business Intelligence***
4. Dicionário de dados

Sistemas OLTP



Data warehouses são, comumente, alimentados por sistemas OLTP independentes.

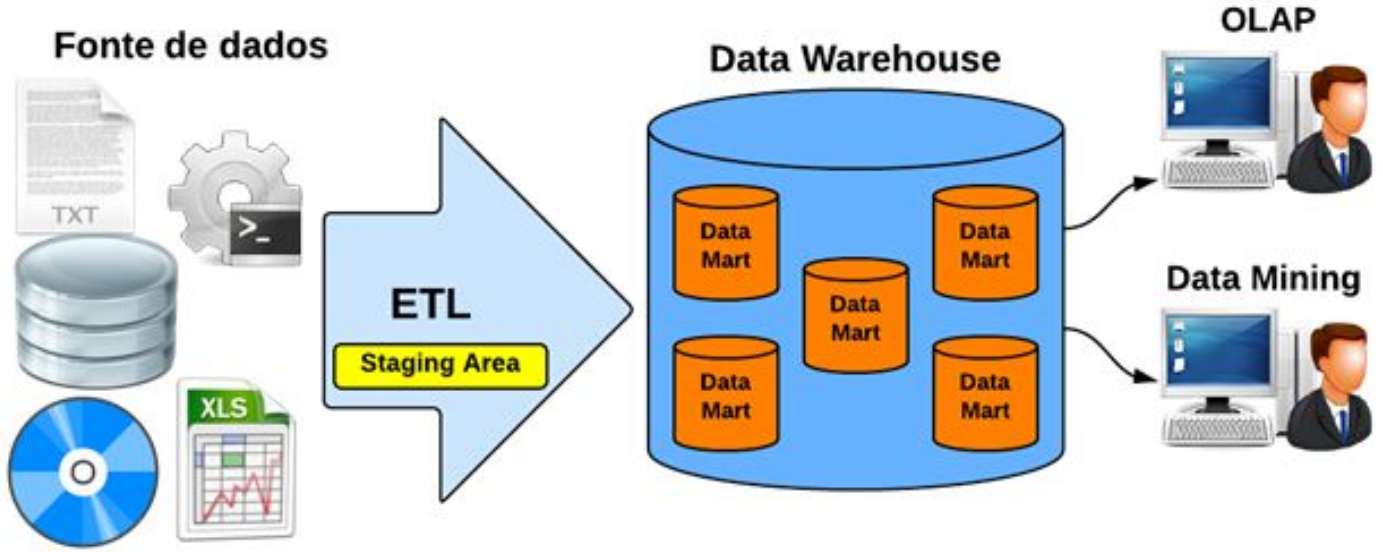
Sistemas OLTP



Data warehouses são, comumente, alimentados por sistemas OLTP independentes.

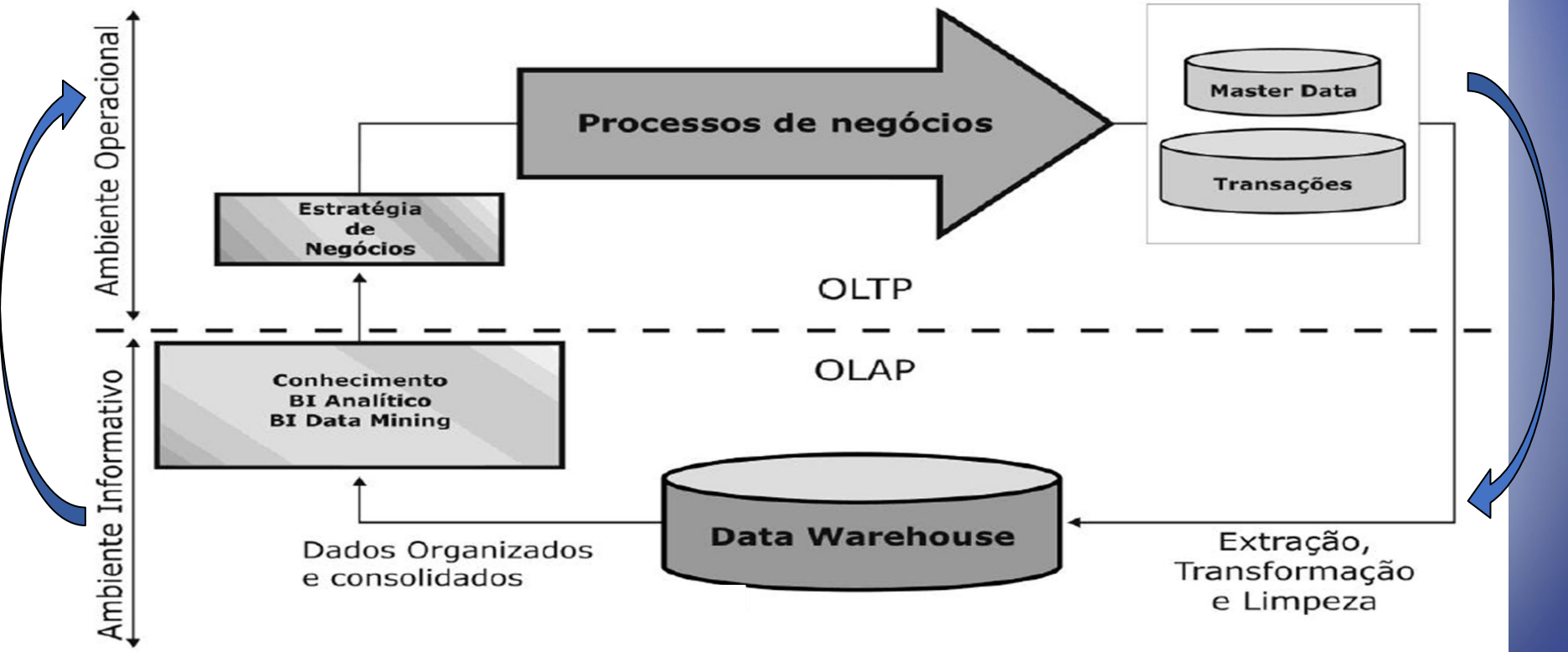
- Sistemas OLTP (***Online Transaction Processing***):
  - toda vez que você vai ao mercado, ao banco ou faz uma compra online, interage em uma rede, usa o GPS, você está usando um sistema OLTP

ETL → DW → OLAP



- Uma vez que os dados foram carregadas via ETL, procede-se com operações analíticas;
- Em se tratando de datawarehousing, técnicas de **Online Analytical Processing** fornecem análises básicas.

Relação OLTP e OLAP



Terminologia

☐ Os termos Data Warehouse, OLTP e OLAP não se referem **apenas a software**;

☐ São termos que englobam software e serviços;

☐ Foram cunhados para a **comunidade empresarial** não possuindo uma correspondência simples e precisa em Ciência da Computação.