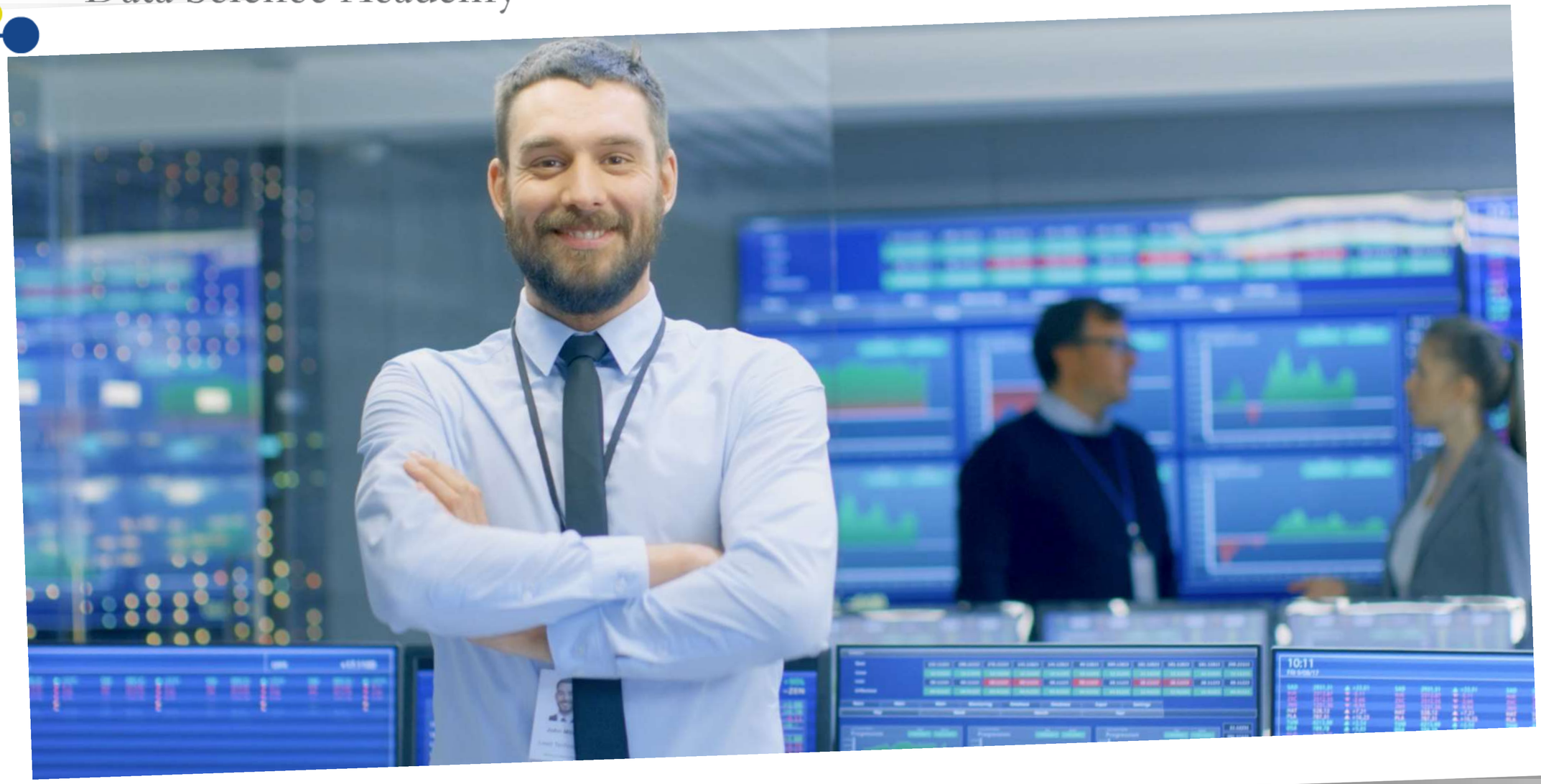
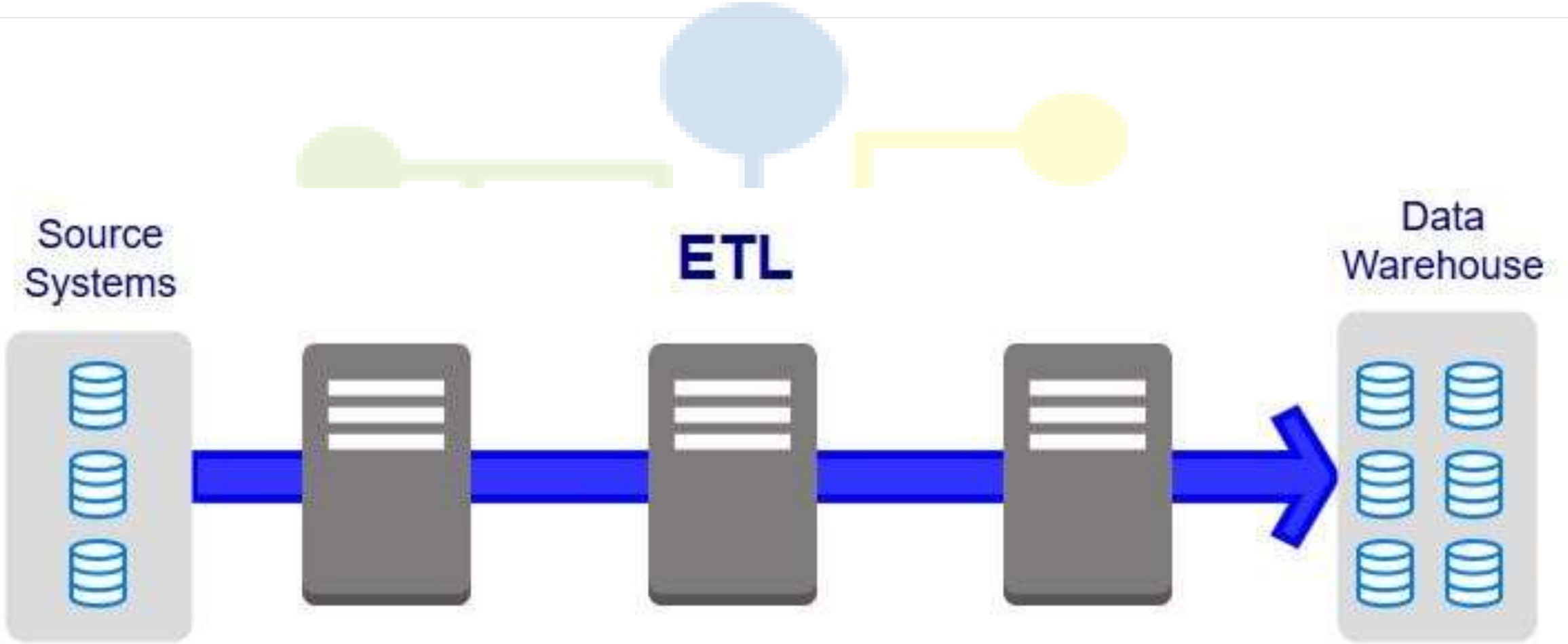
Data Science Academy

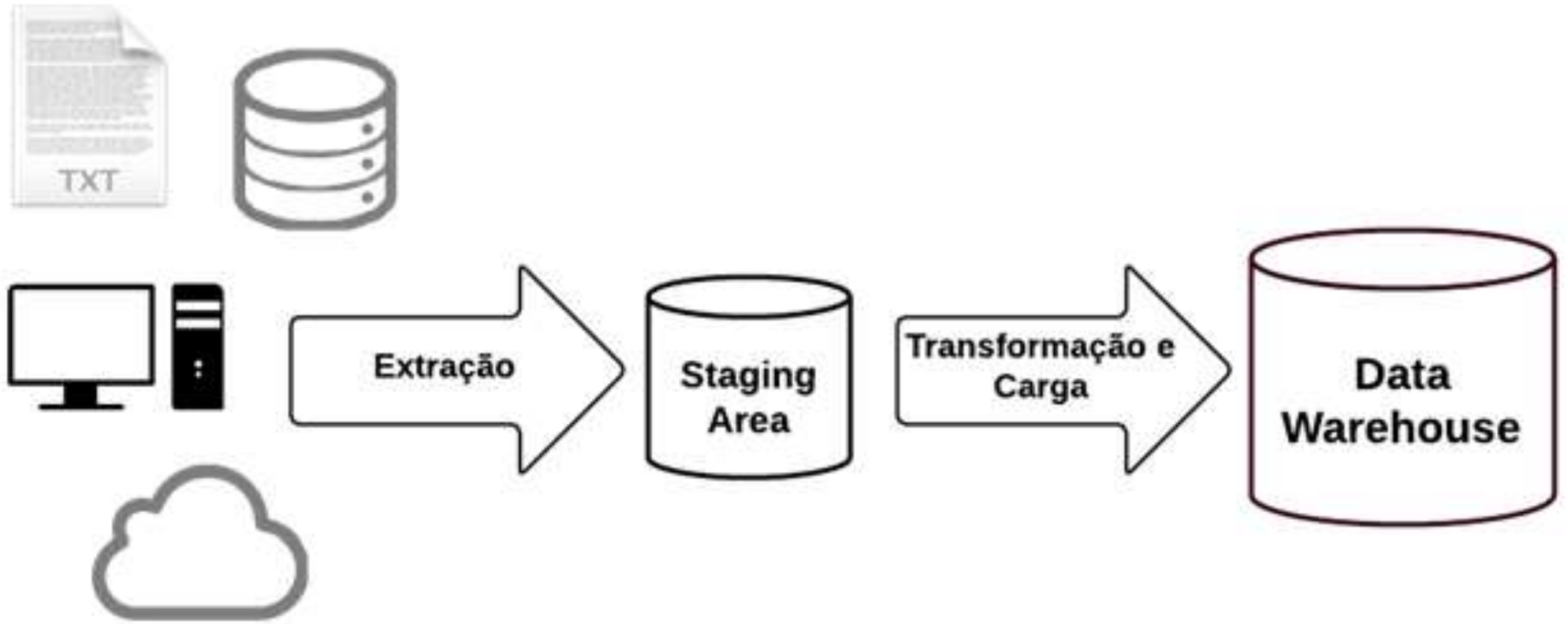


Design e Implementação de Data Warehouses

Data Science Academy

Extração, Transformação e Carga de Dados (ETL)


Source Systems
ETL
Data Warehouse



Data Science Academy

O que estudaremos neste capítulo?

- Definir o processo de ETL
- Identificar as tarefas em ETL
- Identificar métodos e técnicas de extração
- Identificar métodos e técnicas de transformação
- Identificar métodos e técnicas de carga de dados
- Identificar problemas no processo de ETL
- Listar critérios de seleção das ferramentas
- Atividades práticas

Data Science Academy

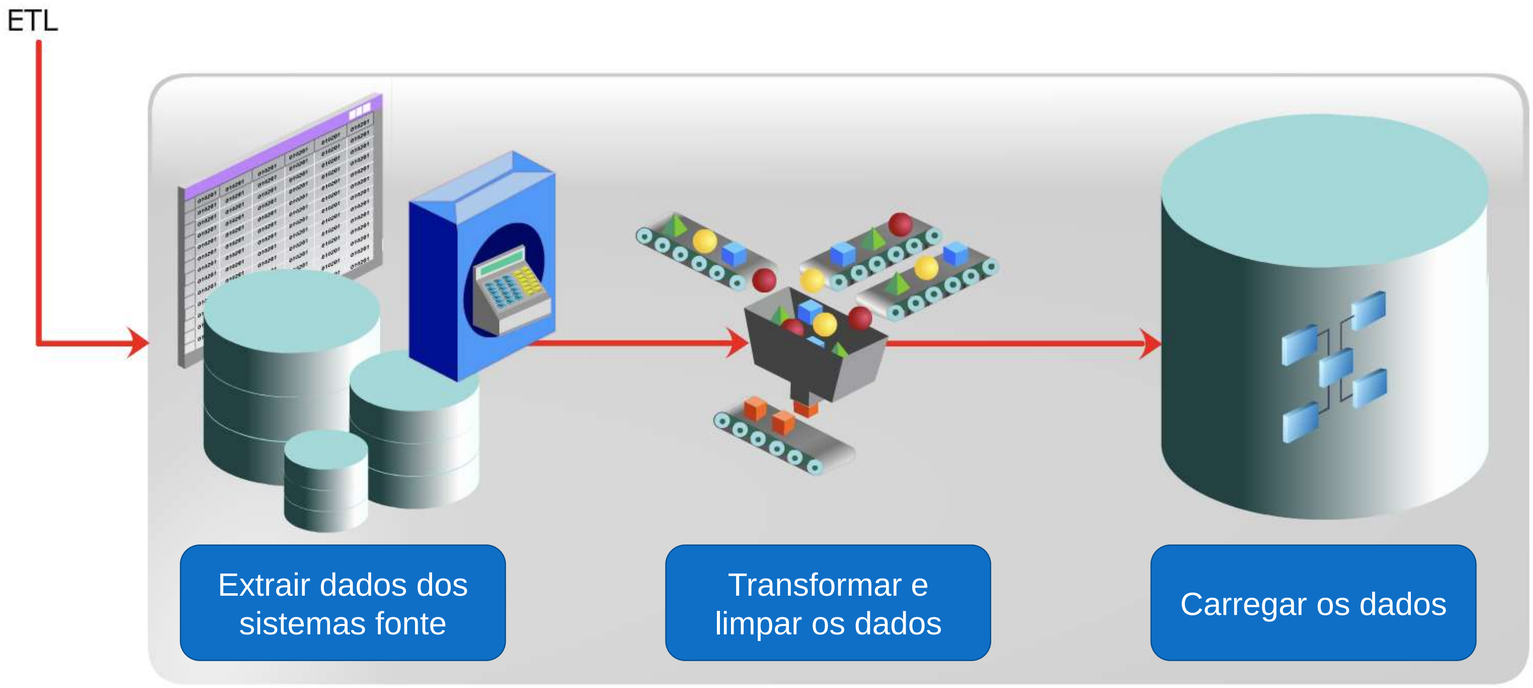
O Que é ETL?

Bancos Relacionais

Sistemas e Ferramentas

Arquivos de Texto

Extract, Transform and Load (Extração, Transformação e Carga)

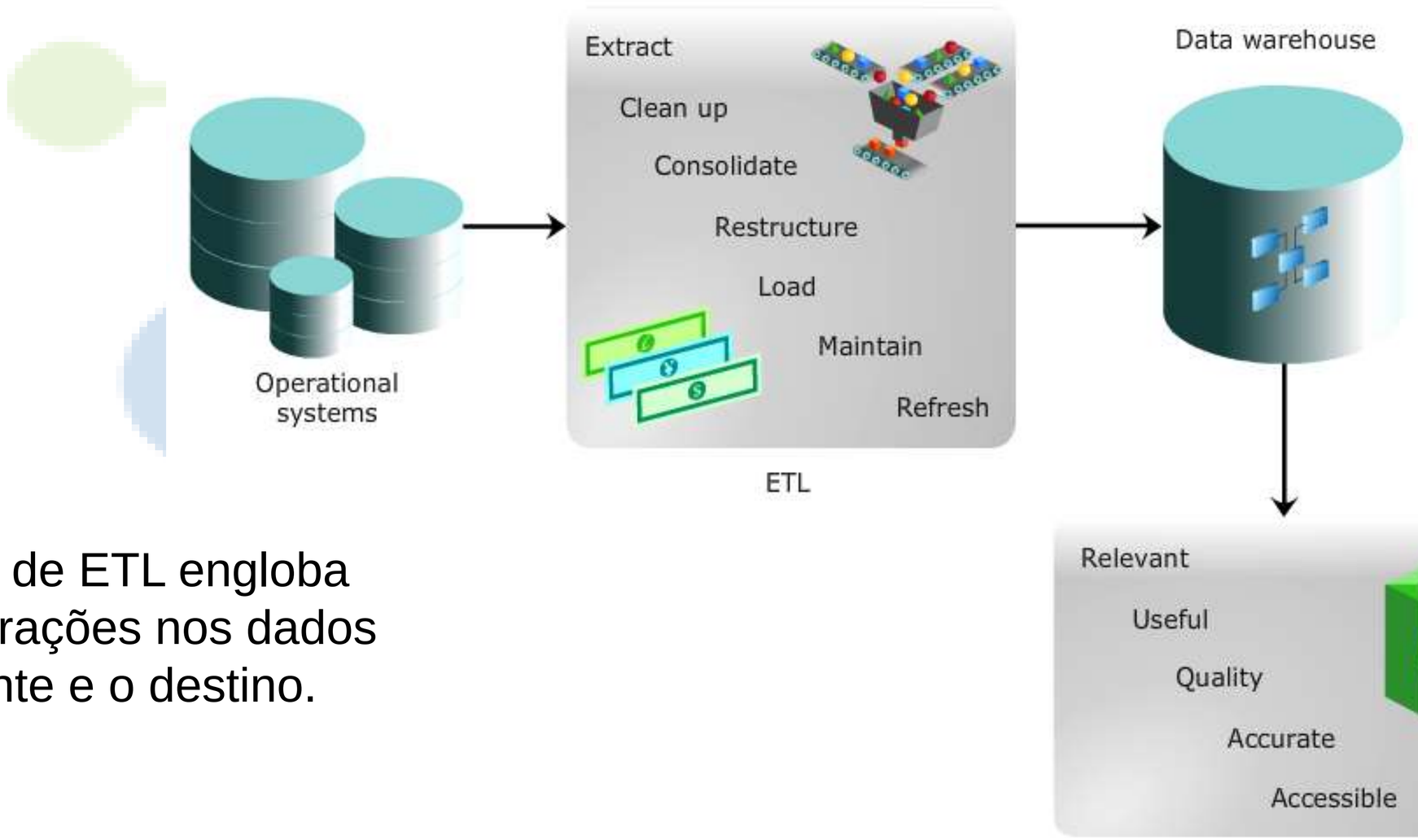


Data Science Academy

O Processo de ETL


Data Science Academy

Data Science Academy

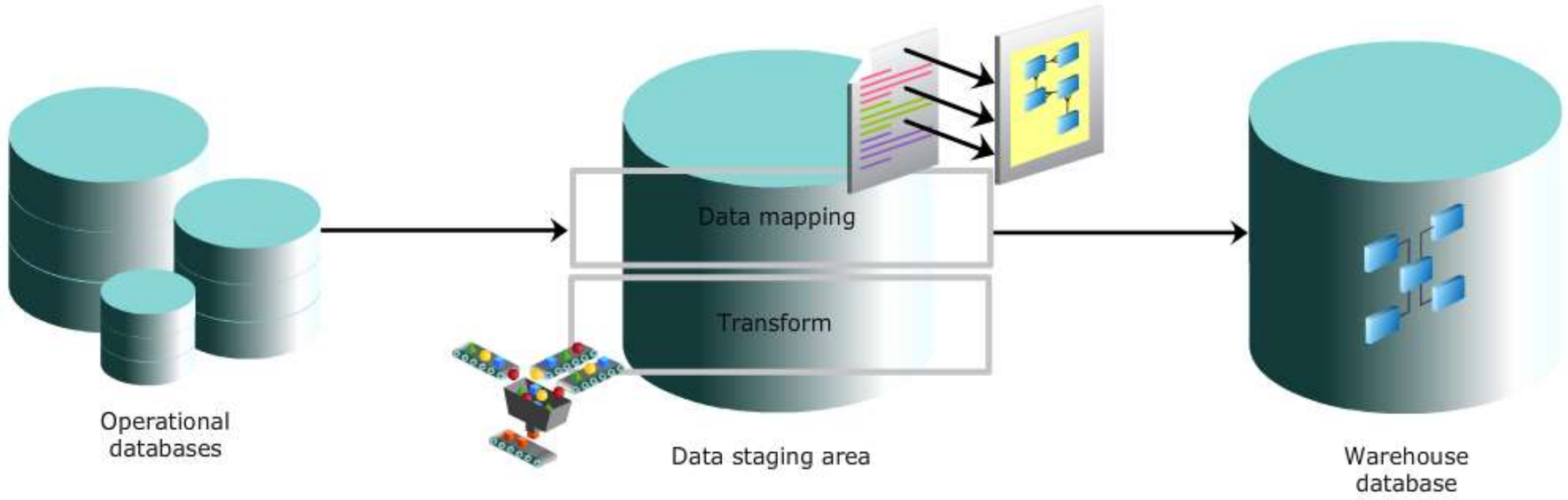


O processo de ETL engloba diversas operações nos dados entre a fonte e o destino.

Data Science Academy

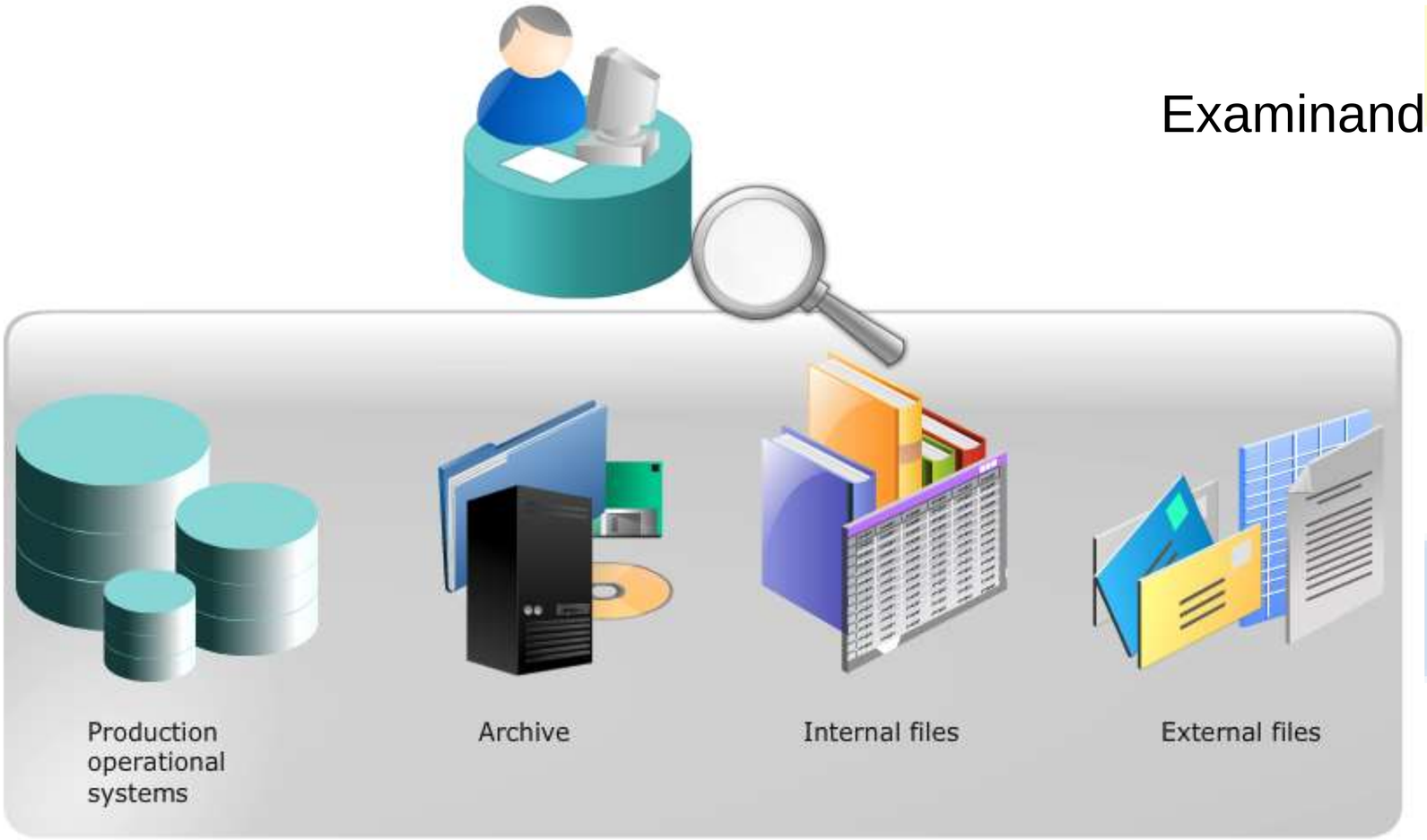
O Processo de Extração

O Processo de Extração

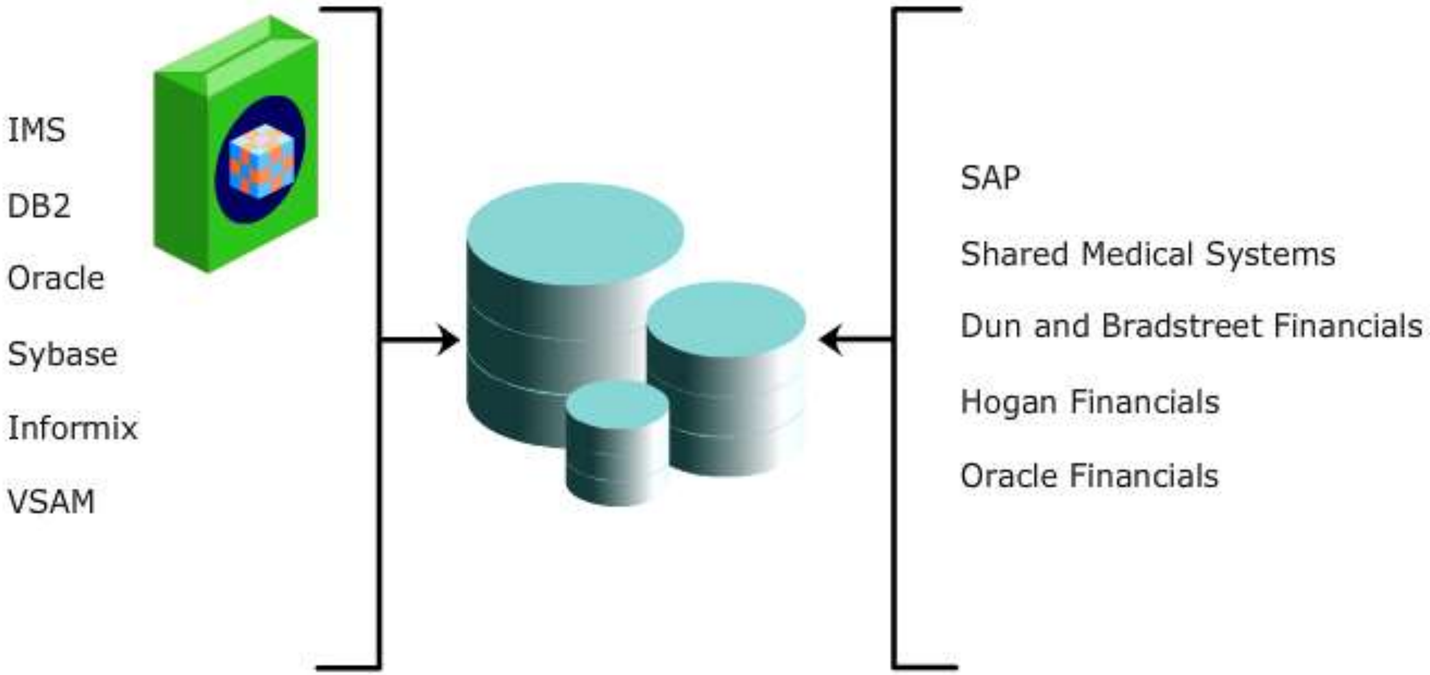


- Dados são extraídos de diversas fontes e em diversos formatos
- Rotinas de extração
  - Regras de negócio
  - Trilhas de auditoria
  - Correção de erros

Examinando as Fontes de Dados

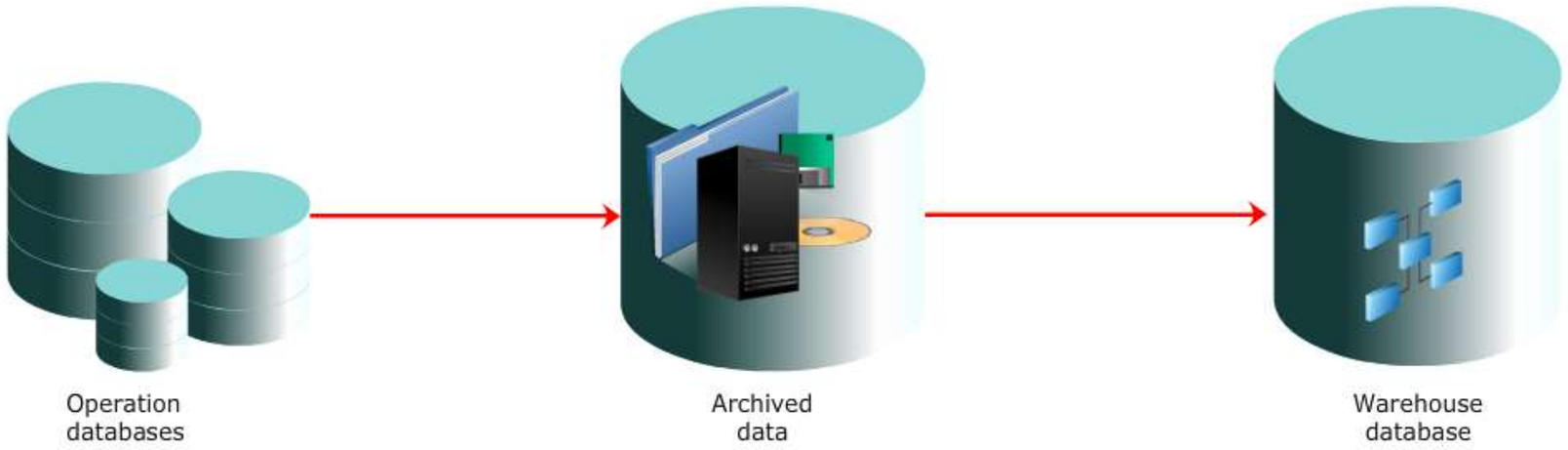


Dados de Produção

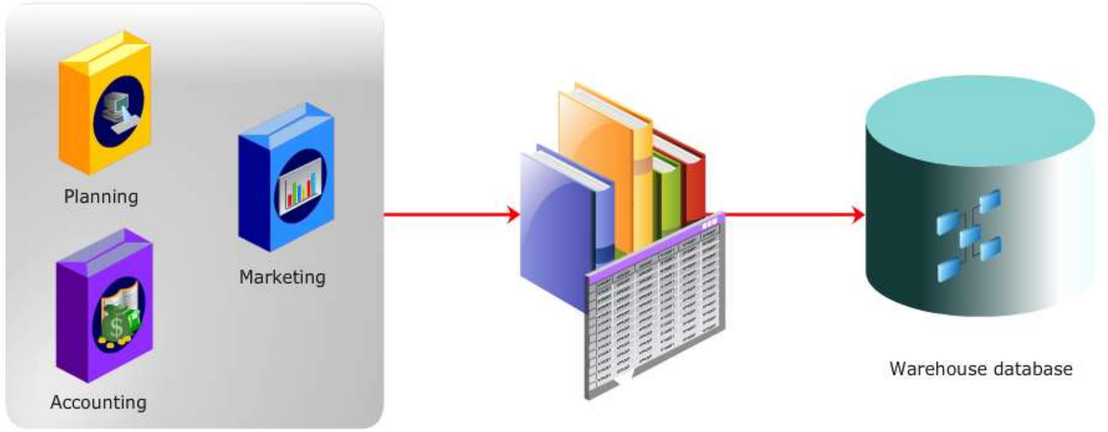


Dados Históricos (Archived Data)

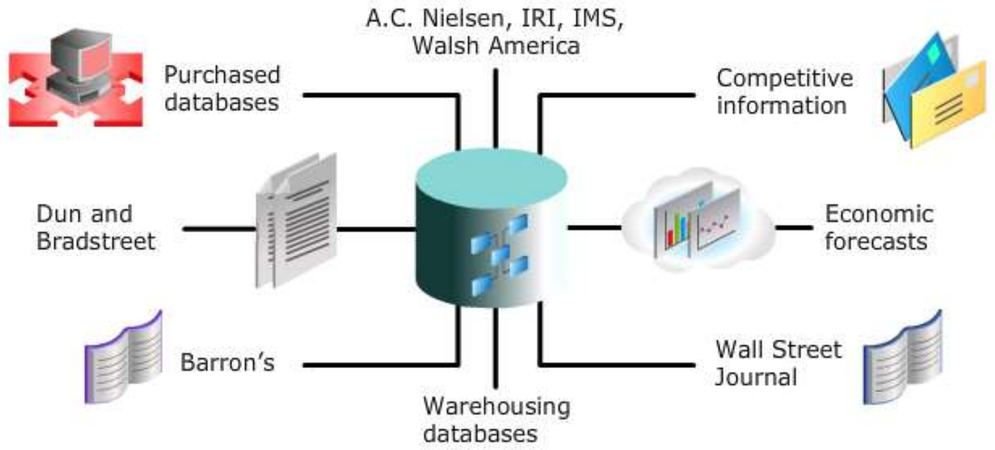
- Úteis para análises ao longo do tempo
- Úteis para a carga inicial
- Podem requerer transformações adicionais



Dados Internos



Dados Externos

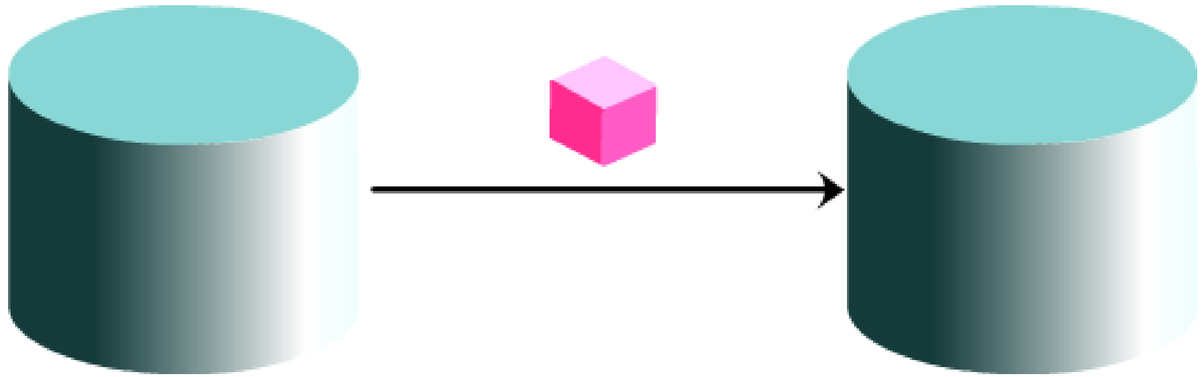


Métodos de Extração

- Extração Lógica
  - Extração Full (Completa)
  - Extração Incremental

- Extração Física
  - Extração Online
  - Extração Offiline

Sua escolha na extração lógica impacta a forma como os dados serão extraídos fisicamente.

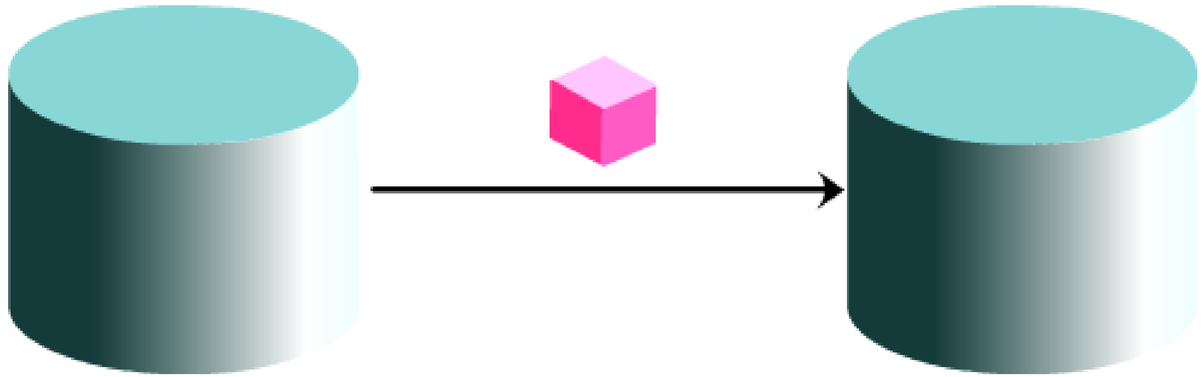


Design do Processo de Extração

- Análise de fontes e tecnologias
- Tipos de dados, qualidade e proprietários
- Refresh: Full ou incremental
- Volume e consistência dos dados
- Automação
- Mantendo os dados extraídos
  - Métodos de acesso
  - Área temporária
  - Tratamento de falhas

Técnicas e Ferramentas de Extração e Manutenção de Metadados

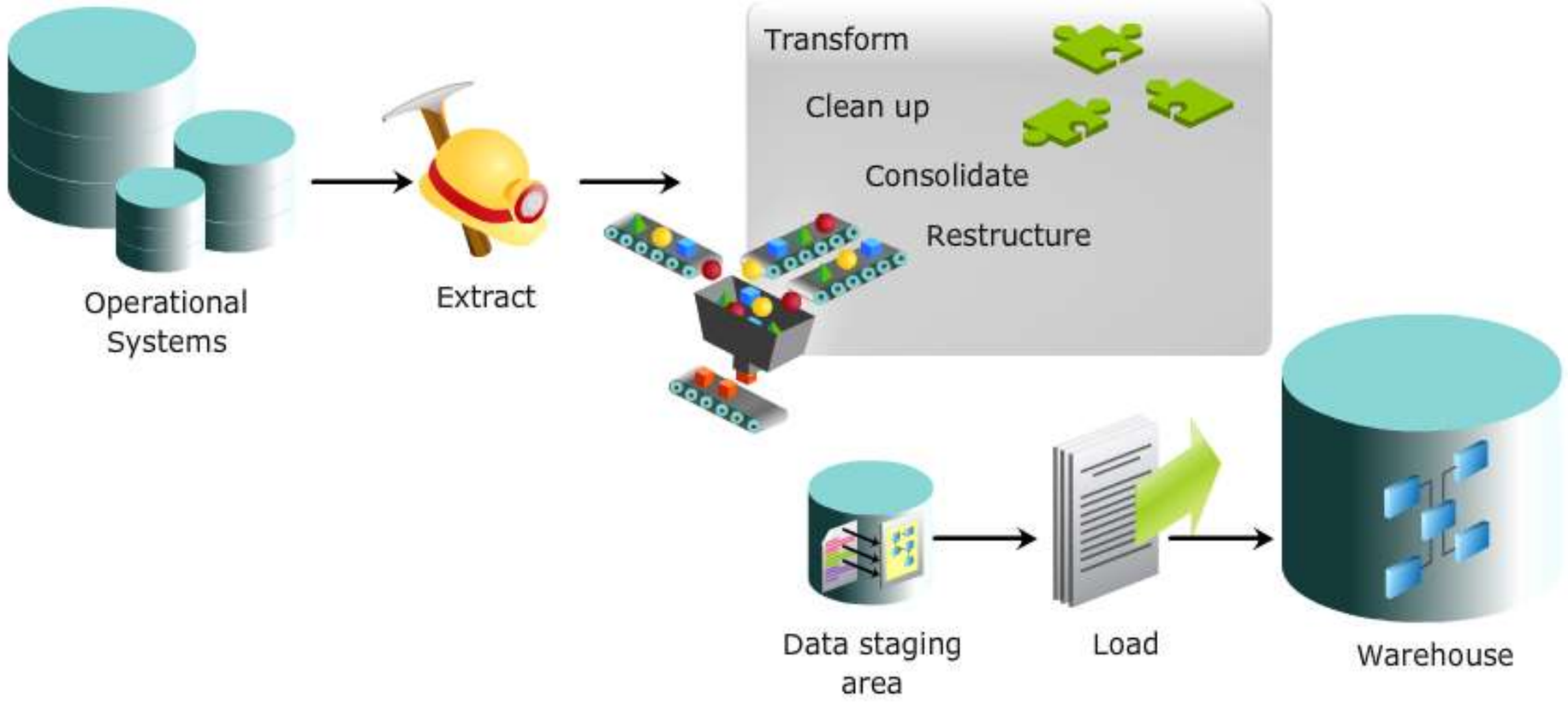
- Programas em C, C++, Python, R, Java ou PL/SQL
- Ferramentas ETL: OWB, ODI, PowerCenter, DataStage
- Ferramentas Oracle: Data Pump, SQL*Loader, External Tables, Oracle Golden Gate



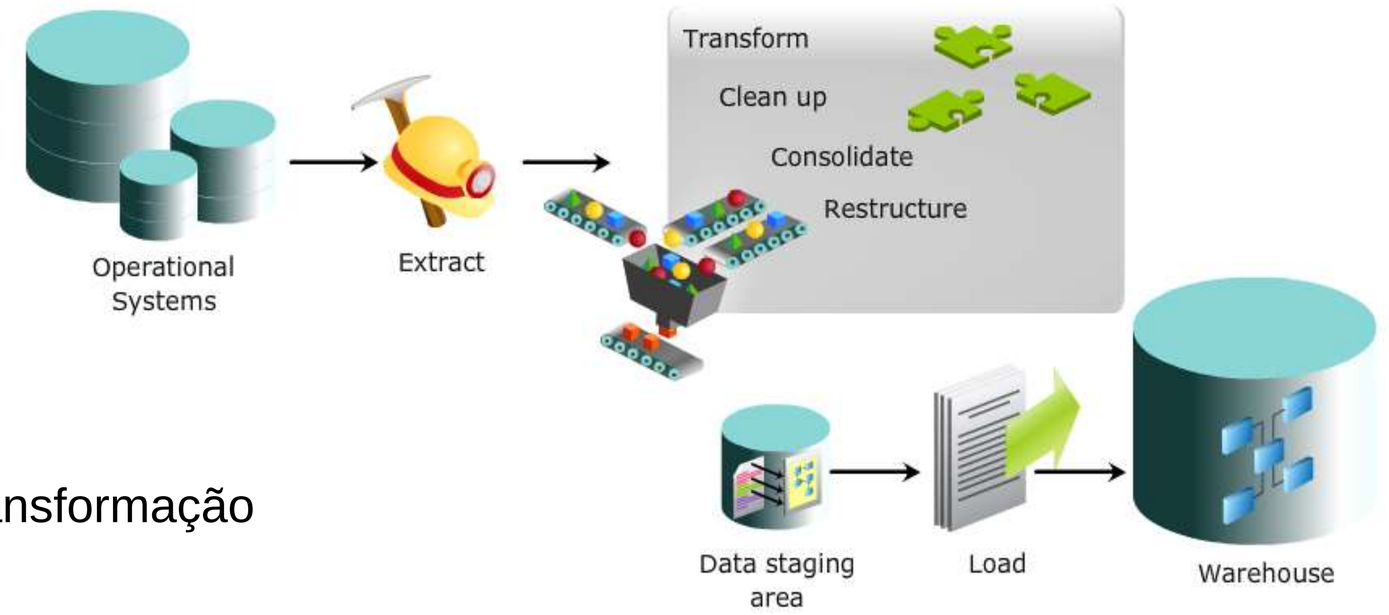
Data Science Academy

O Processo de Transformação

O Processo de Transformação



O Processo de Transformação



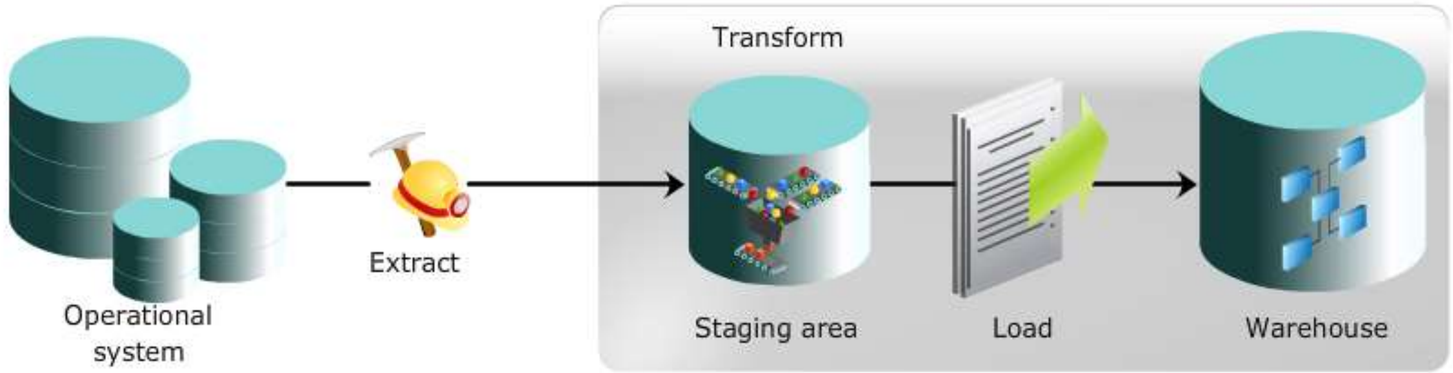
Principais tarefas no processo de transformação

- Definir as transformações
- Identificar modelos de Staging Area
- Identificar e eliminar anomalia nos dados
- Design de transformação de dados

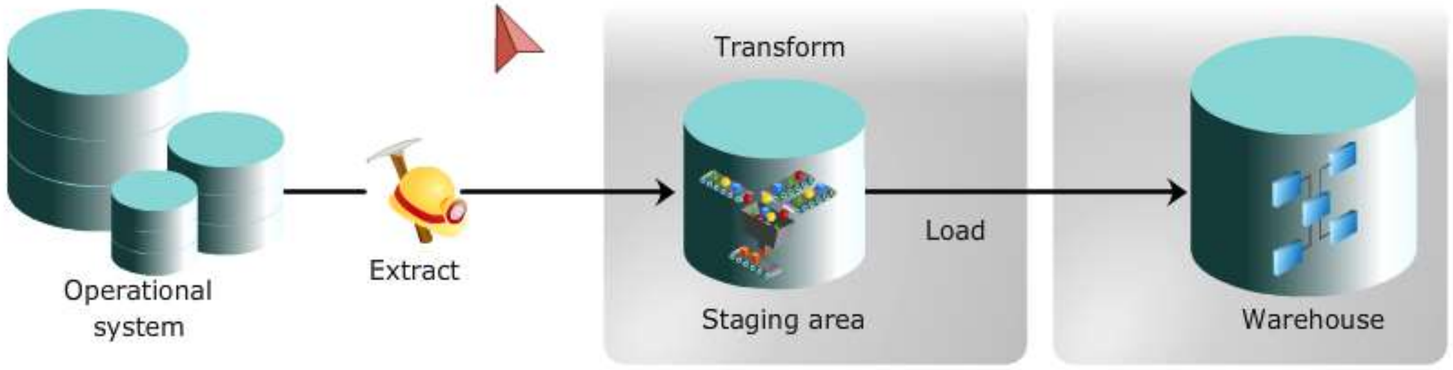
Modelos de Staging
Remoto e On-site

Modelo de Staging Remoto

Data Staging fica dentro do ambiente do DW

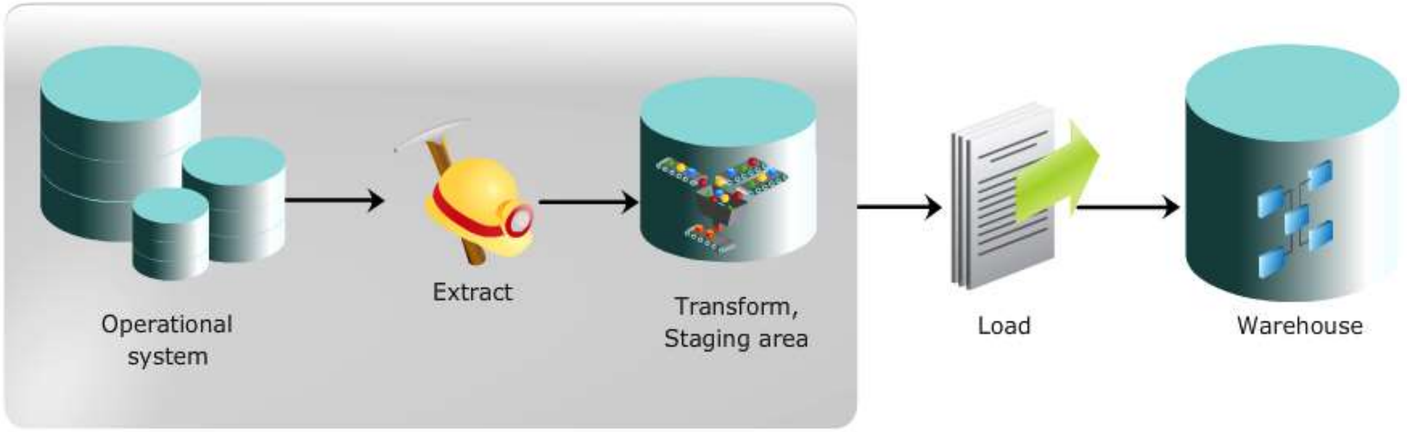


Data Staging tem seu próprio ambiente



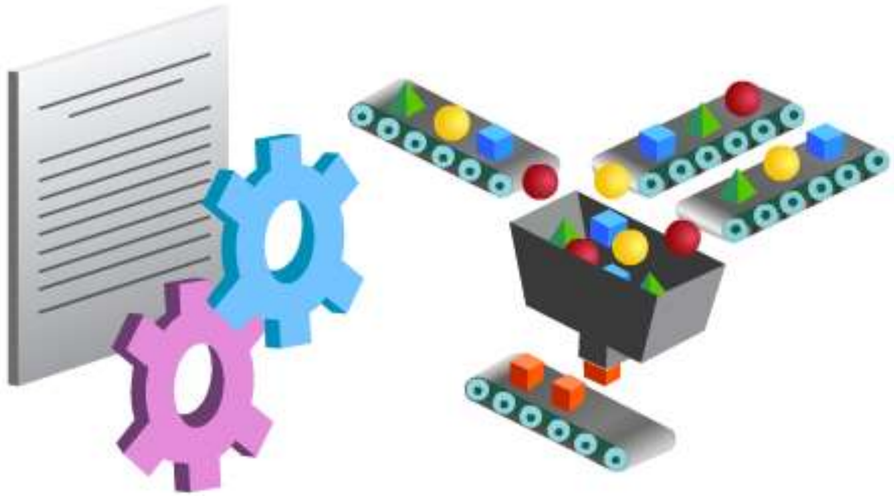
Modelo de Staging On-site

Data Staging fica dentro do ambiente operacional (fonte)



Rotinas de Transformação

- Limpeza de Dados
- Eliminação de inconsistências
- Inclusão de elementos
- Merging de dados
- Integração



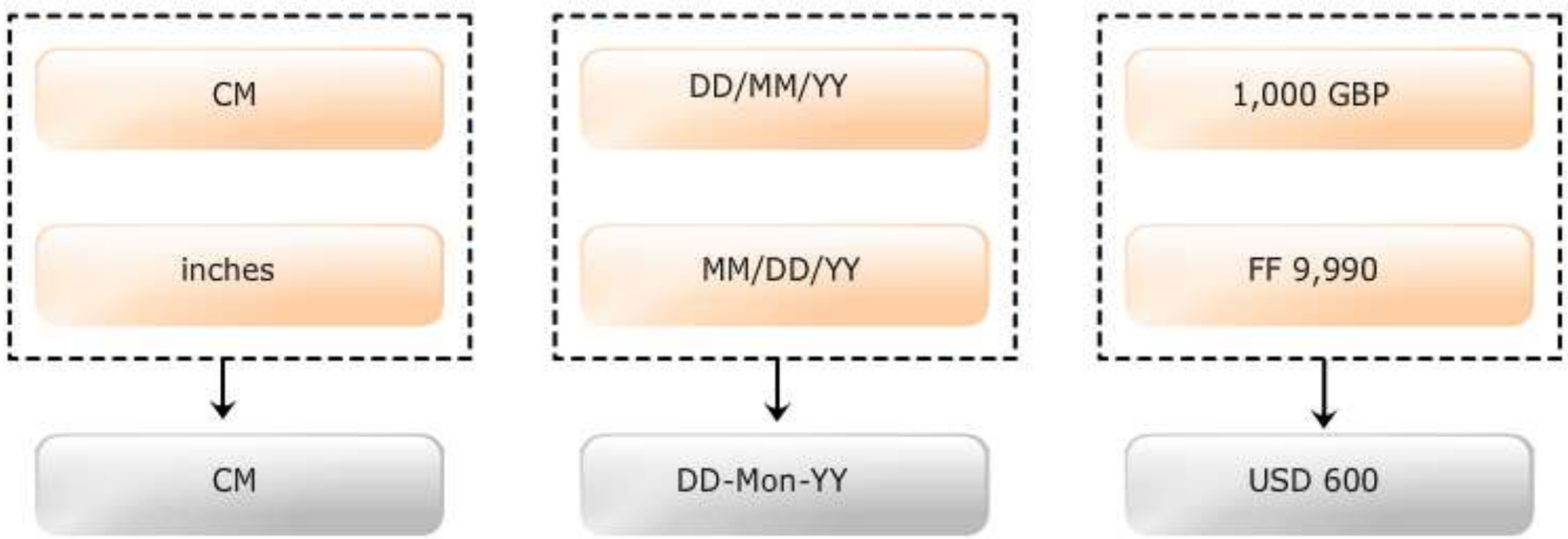
Transformação – Problemas e Soluções

Anomalia nos dados devem ser removidas

Codigo_Cliente	Nome_Cliente	Endereço_Cliente
123456	Data Science Academy	Rua Acre 786
123457	DSA	Av Acre 786 – Sala 18
123458	DS Academy	Avenida Acre 786
123459	Data Science Academia	Avenida Acre, 786

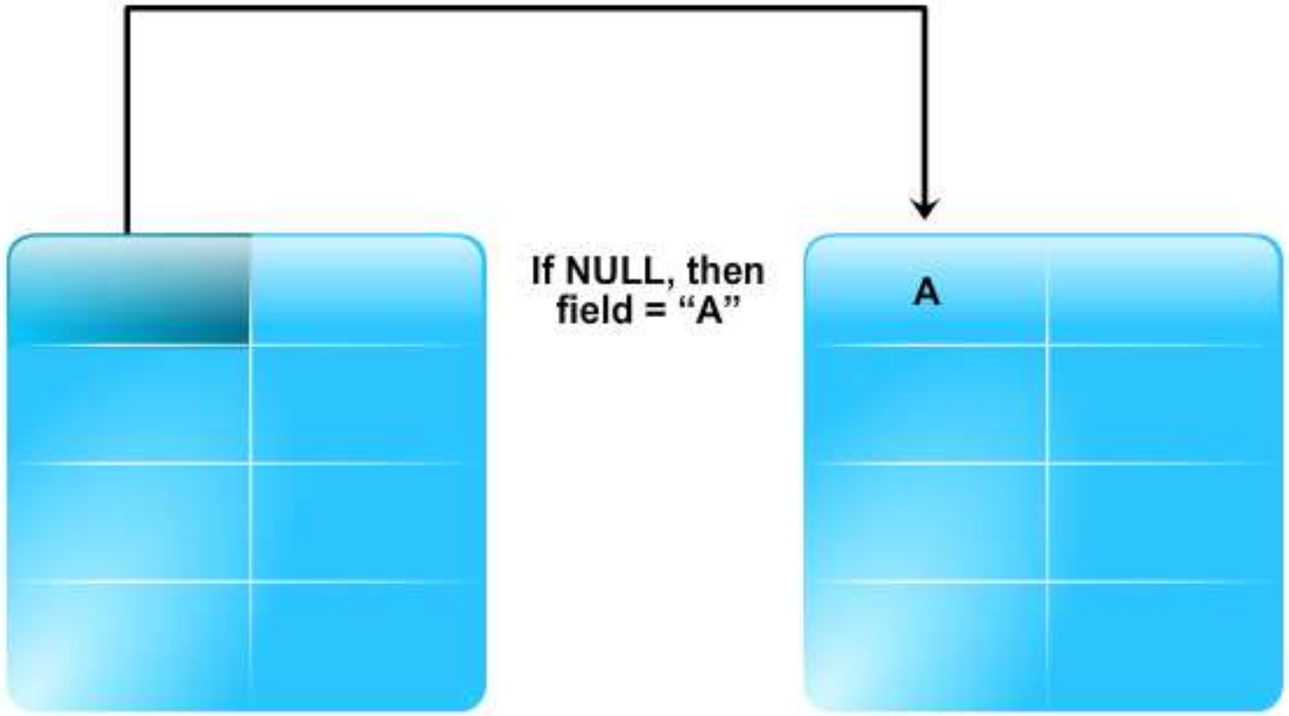
Transformação – Problemas e Soluções

Padronização



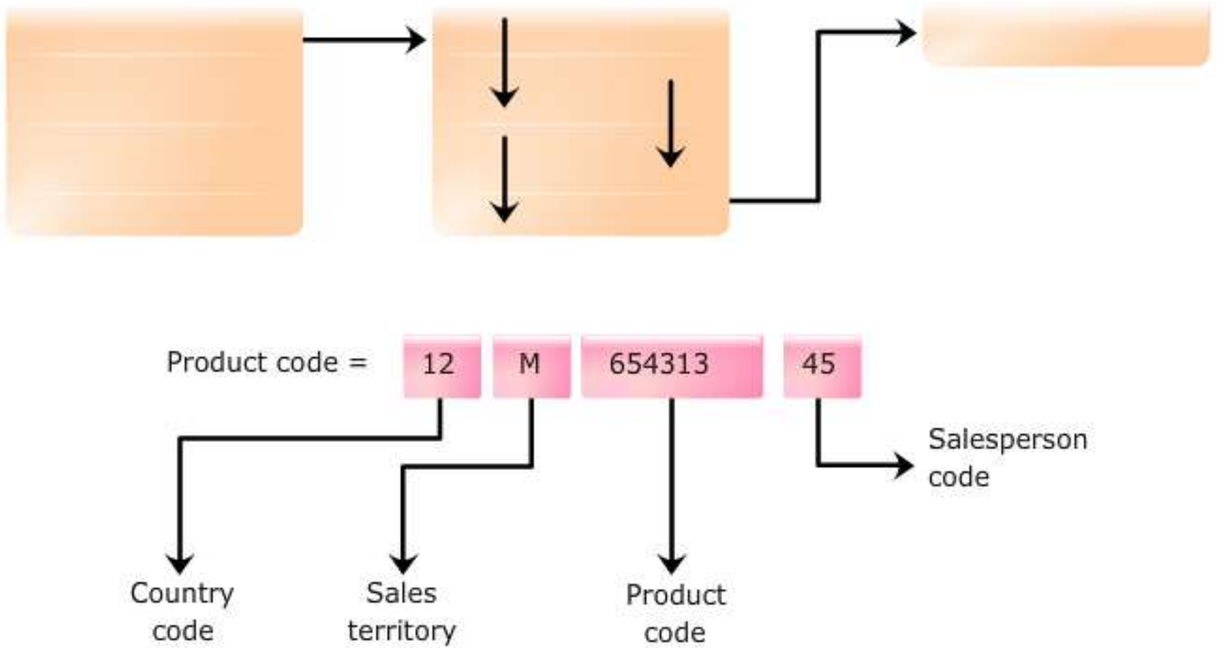
Transformação – Problemas e Soluções

Valores Missing



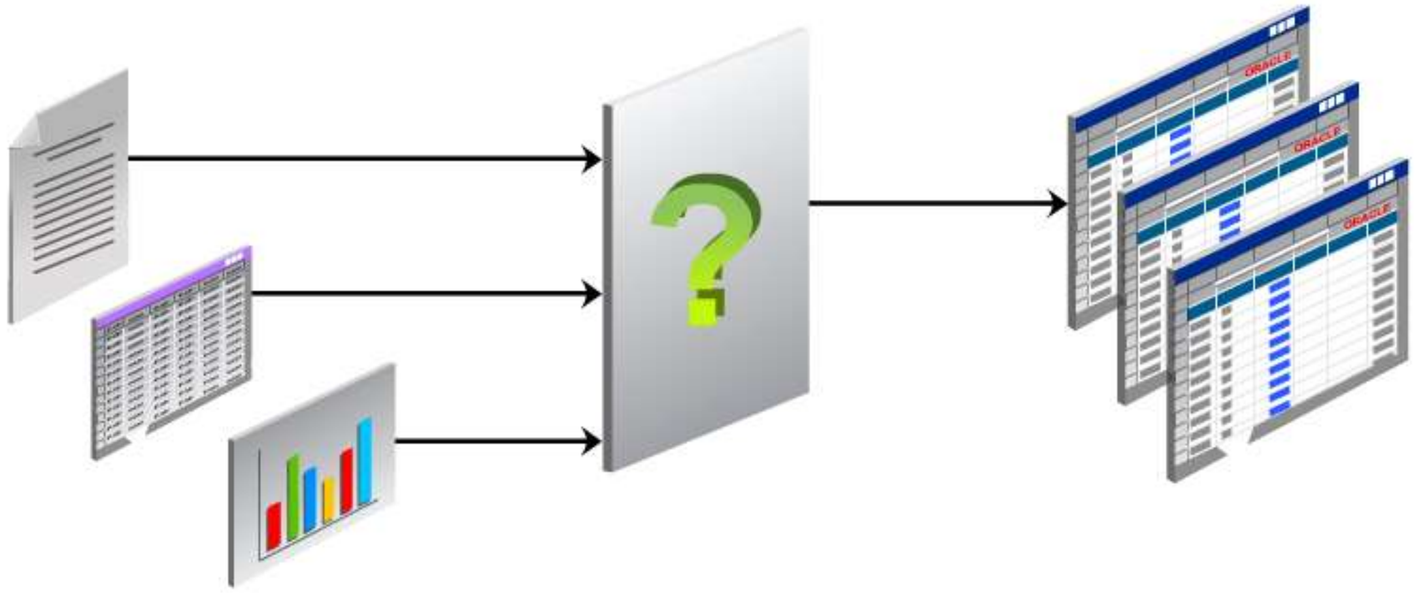
Transformação – Problemas e Soluções

Multipart Keys



Transformação – Problemas e Soluções

Integração de Múltiplas Fontes de Dados



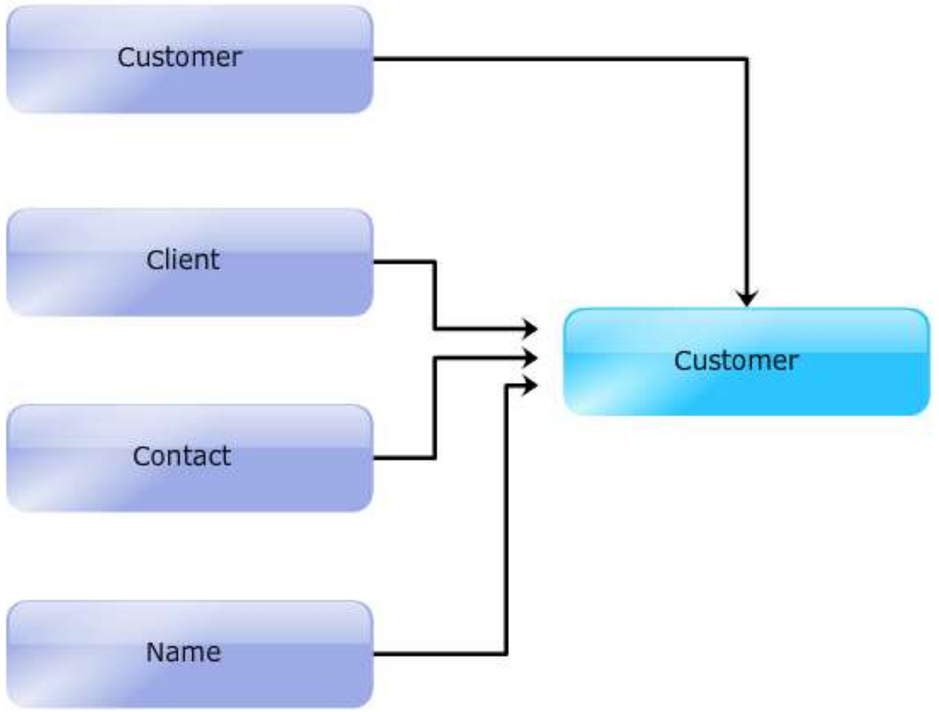
Transformação – Problemas e Soluções

Valores Duplicados

```
SELECT ...
FROM table_a, table_b
WHERE table_a.key (+)= table_b.key
UNION
SELECT ...
FROM table_a, table_b
WHERE table_a.key = table_b.key(+);
```

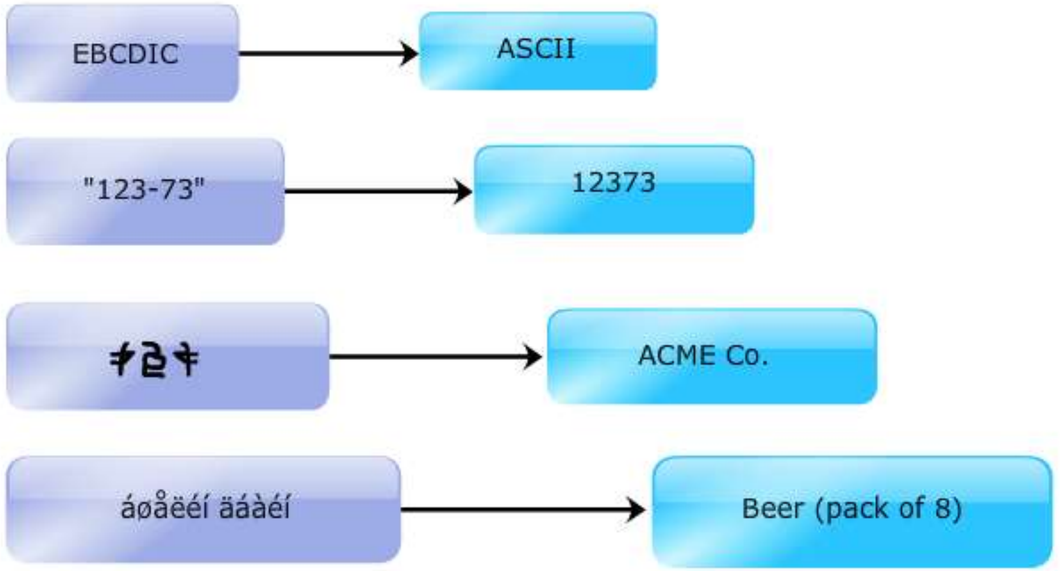
Transformação – Problemas e Soluções

Convenção de Nomes



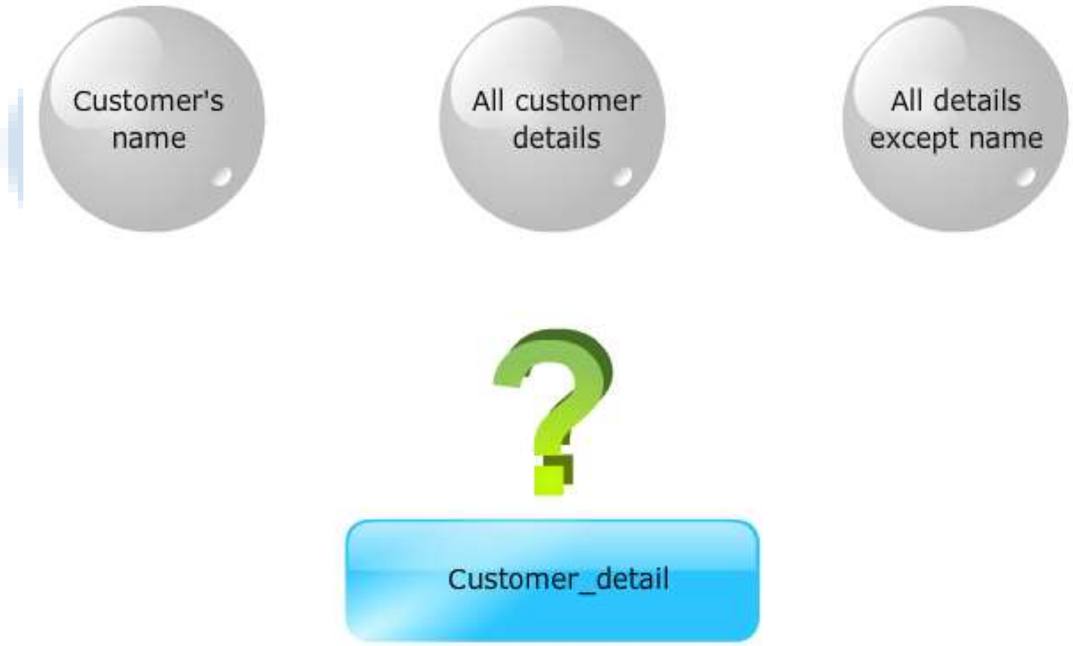
Transformação – Problemas e Soluções

Formatos dos Dados



Transformação – Problemas e Soluções

Significado de Elementos



Transformação – Problemas e Soluções

Integridade Referencial

Department
10
20
30
40

Emp	Name	Department
1099	Smith	10
1289	Jones	20
1234	Doe	50
6786	Harris	60

Padrões de Qualidade dos Dados

- Definir estratégias de qualidade
- Decidir o nível de qualidade aceitável
- Qualidade deve ser uma preocupação também na fonte de dados
- Considere sugerir alterações na fonte
- Design do processo de limpeza
- Limpeza inicial e processo de limpeza no Refresh podem ser diferentes

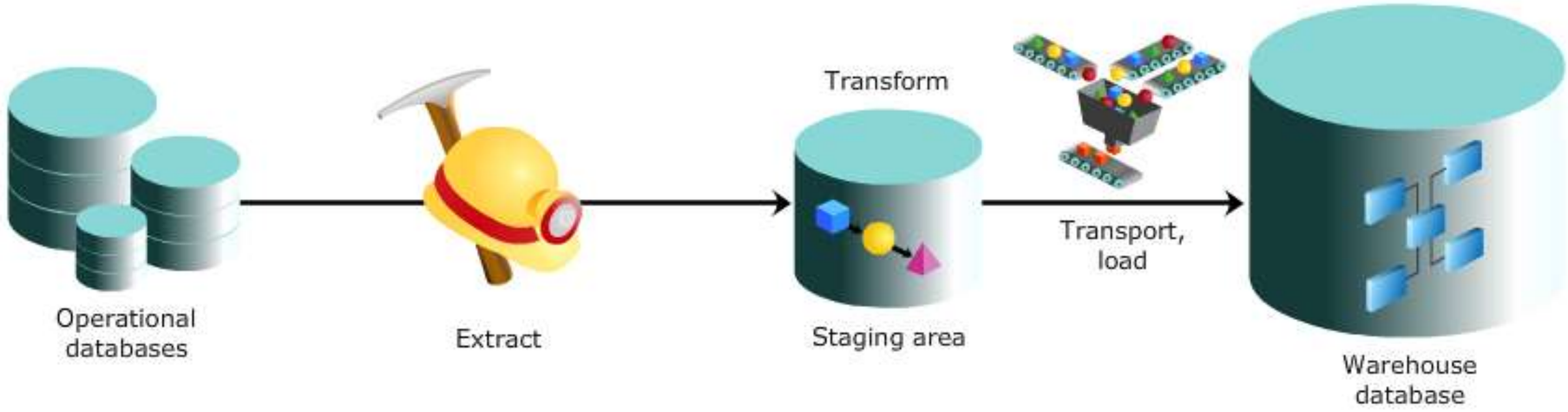
Design do Processo de Transformação

- Dados operacionais NÃO devem ser carregados diretamente no DW
- Os dados devem ser limpos e tratados (não suponha que já estão)
- Os dados na fonte podem variar entre as operações de Refresh
- Defina uma estratégia de Staging Area
- Documente todo o processo (o documento deve ser atualizado a cada alteração)
- Verifique se a empresa possui um padrão de nomenclatura e de limpeza
- Assuma a responsabilidade
- Resolva problemas
- Um Analista de Qualidade pode ser útil durante esta etapa do ETL

Data Science Academy

O Processo de Carga dos Dados

O Processo de Carga dos Dados carrega os dados no DW, após as etapas de extração e transformação.



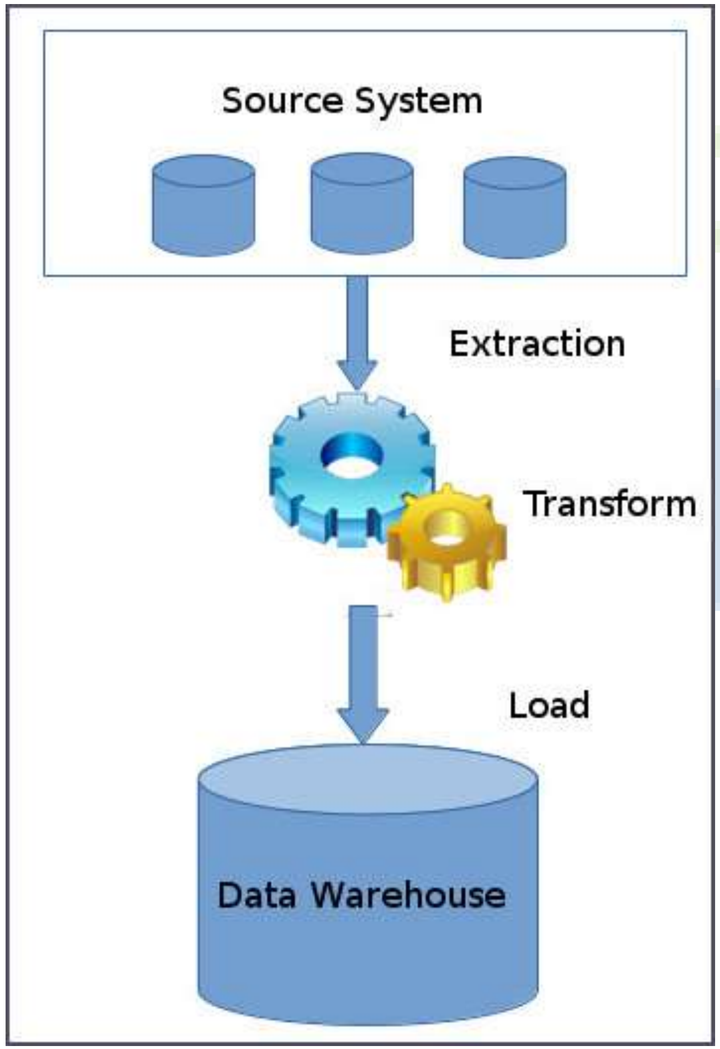
Como transportar os dados da Staging Area (após o processo de transformação), para o DW?

Flat Files

Distributed Systems

Transportable Tablespaces

O Processo de Carga dos Dados



Carga Inicial

- A carga inicial de dados é um evento único responsável por popular o DW com os dados disponíveis até o momento da carga.
- Envolve grandes quantidades de dados.
- Grande quantidade de processamento e uso de recursos computacionais.

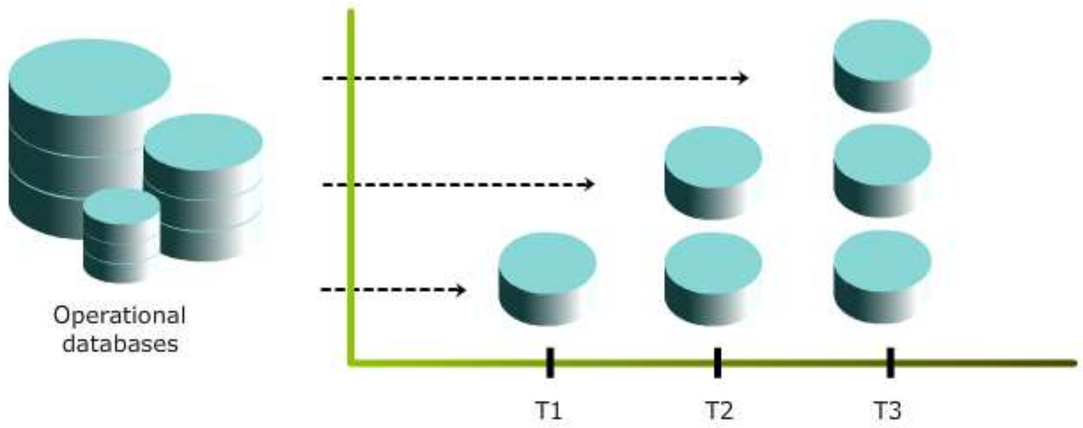
Refresh

- Realizado de acordo com o ciclo de negócio.
- Volume menor de dados a ser carregado.
- Menor processamento.

Modelos de Refresh de Dados

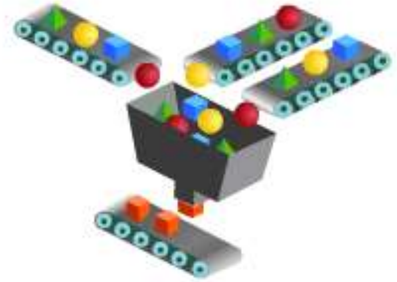
- **Extract Processing Environment** (Em intervalos específicos, cria um novo snapshot na fonte de dados para carregar no DW).

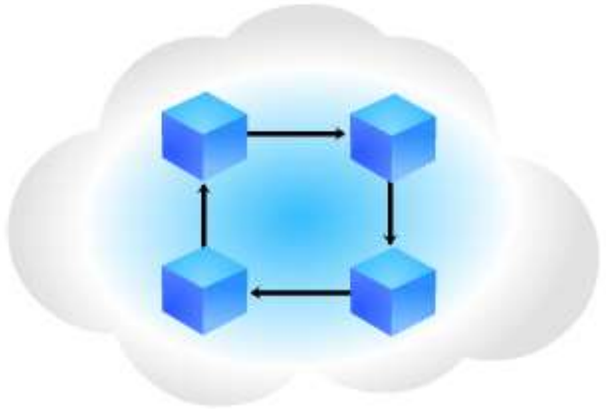
- **Warehouse Processing Environment** (A cada intervalo de tempo, carrega no DW, mudanças na fonte de dados).



Construindo o Processo de Carga de Dados

- Ferramentas que serão usadas
- Métodos de transporte dos dados
- Janela de carga dos dados
- Tempo de execução das tarefas pré-carga
- Ciclo de Refresh
- Largura de banda e conectividade
- Testar a solução proposta
- Documentar
- Monitorar, revisar, melhorar

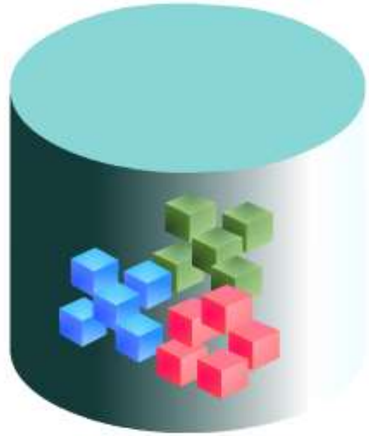




Granularidade dos Dados

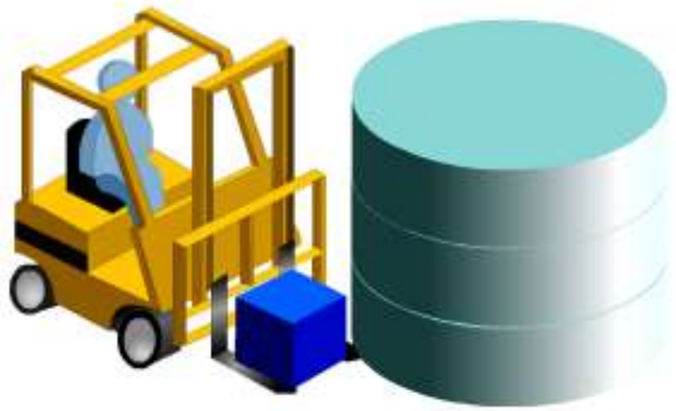
- Alta Granularidade
  - Mais dados, mais espaço em disco, mais tempo de processamento, mais caro, mais detalhes nos relatórios

- Baixa Granularidade
  - Menor volume de dados, menos espaço em disco, menos detalhes nos relatórios



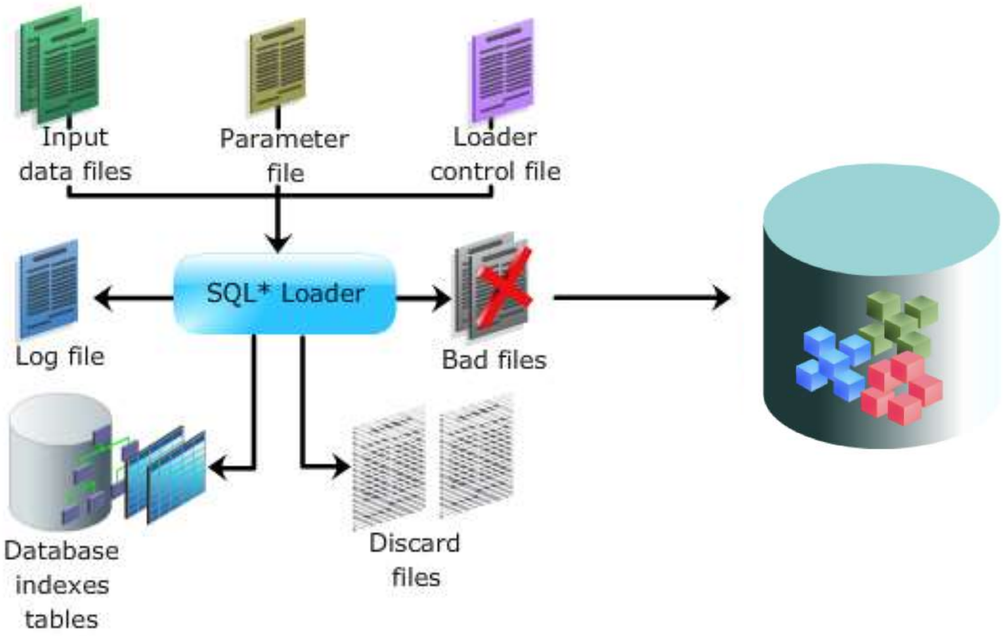
Técnicas de Carga de Dados

- Selecionar uma ferramenta de carga de dados
- Suporte a ambientes distribuídos
- Acesso e execução em tempo real
- Evitar programas muito customizados
- Considerar taxas de transferência



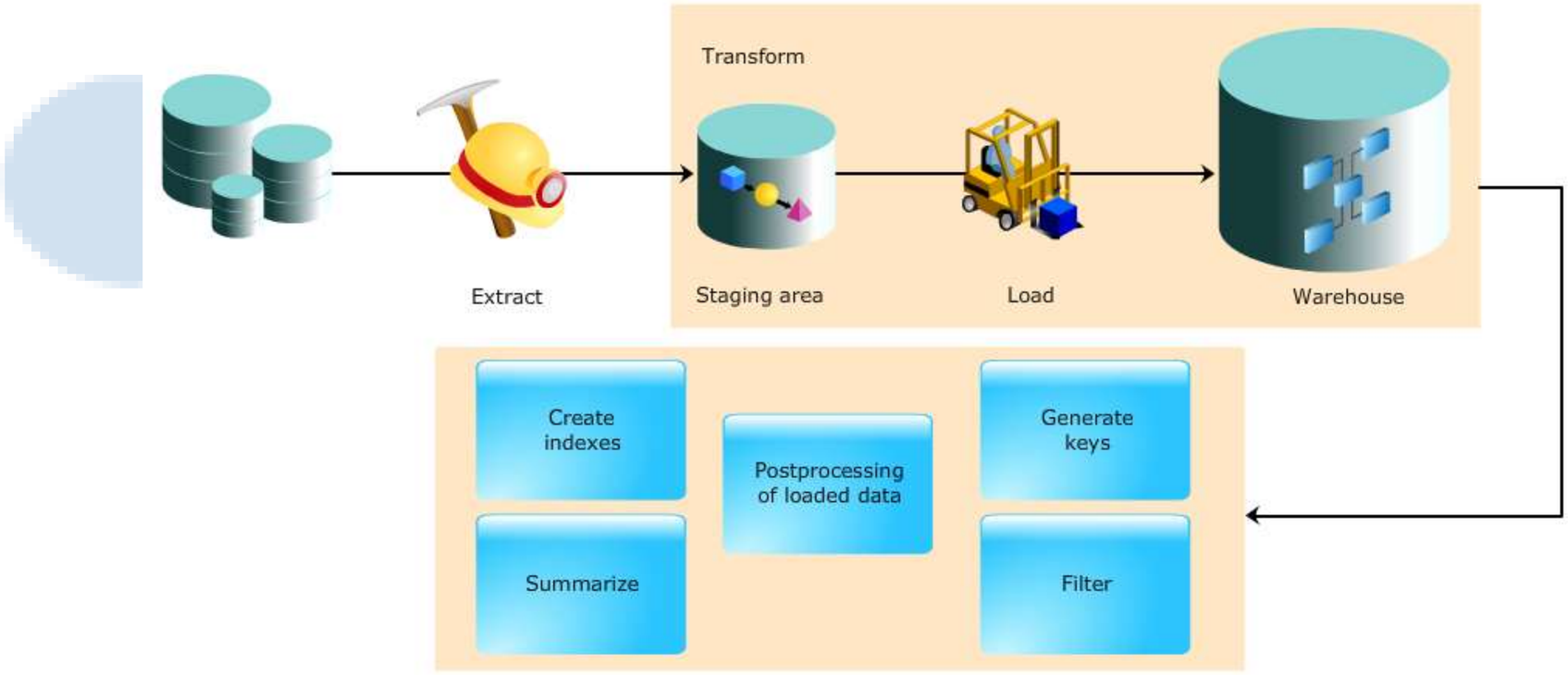
Ferramentas de Carga de Dados

- **SQL*Loader**
- **Data Pump Export/Import**
- **External Tables**
- **ODI**
- **Pentaho**
- **PowerCenter**
- **DataStage**
- **Procedures PL/SQL**

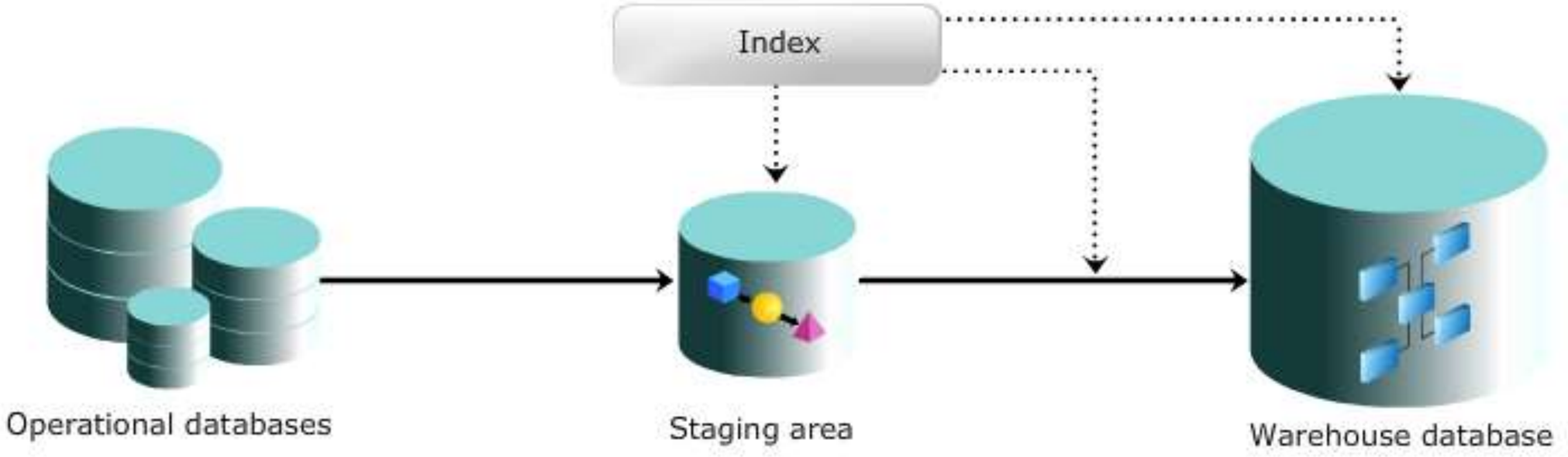


Processo Pós Carga de Dados

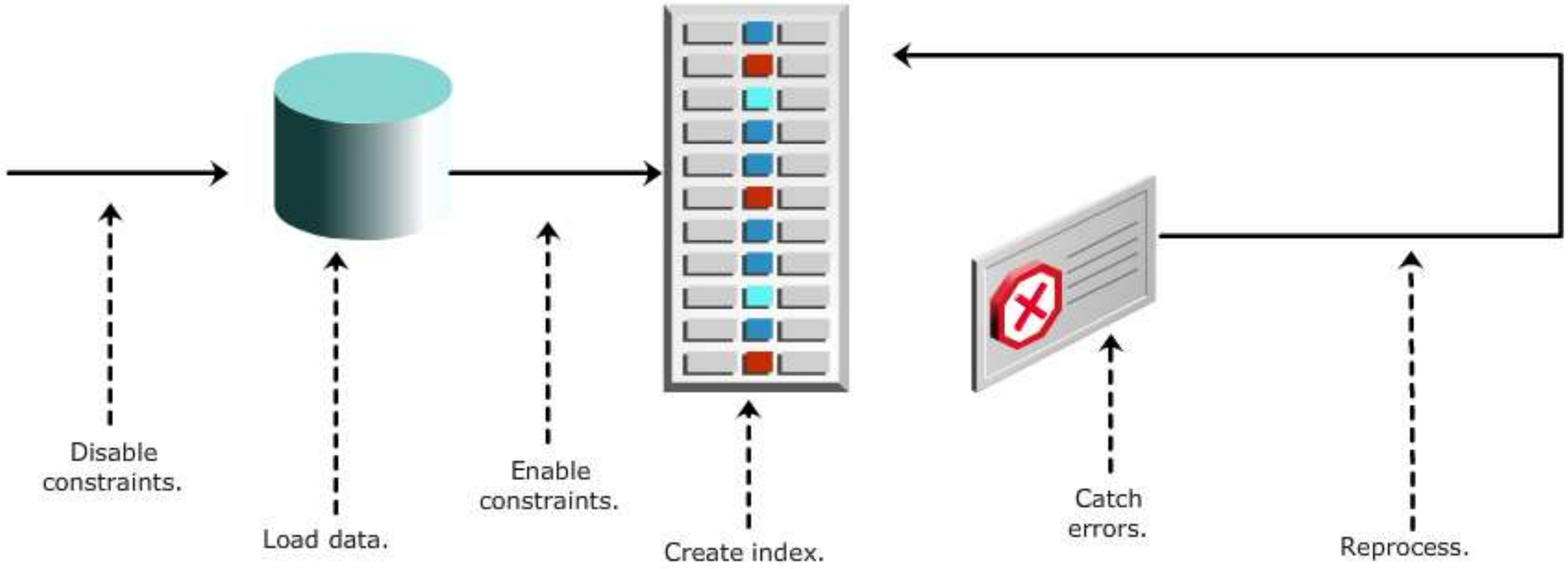
- Criar os índices
- Gerar chaves
- Criar tabelas de sumário
- Filtrar dados



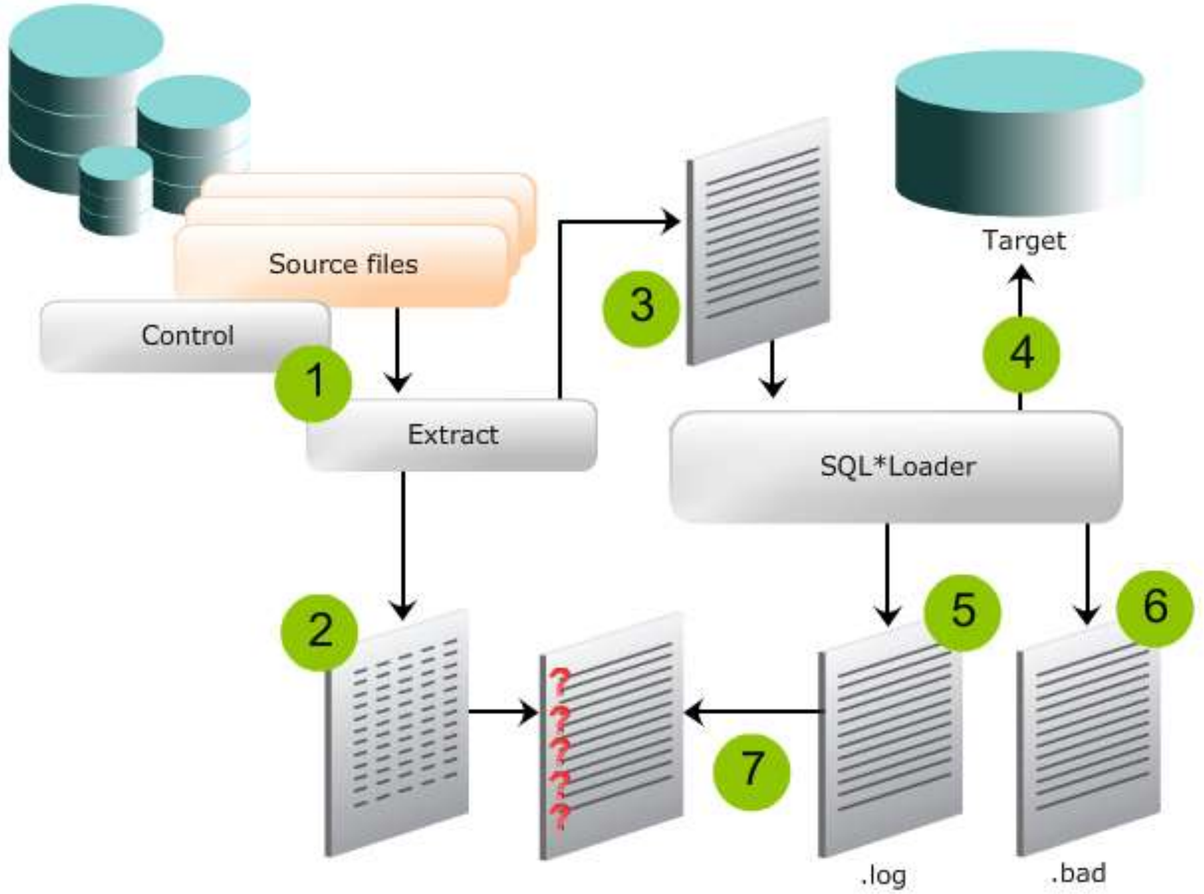
Indexação


Index
Operational databases
Staging area
Warehouse database

Indexação

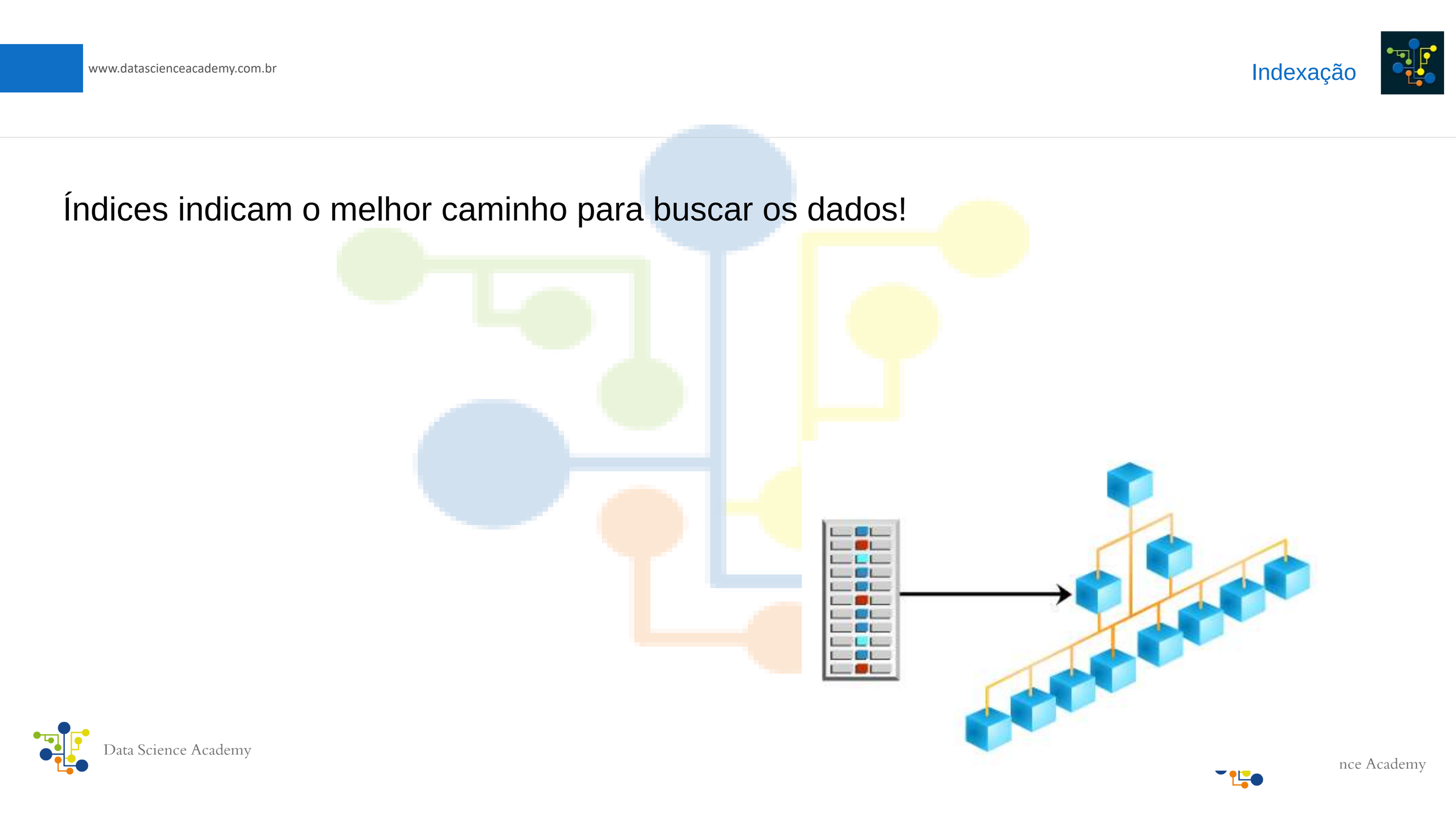

Disable constraints.
Load data.
Enable constraints.
Create index.
Catch errors.
Reprocess.

Verificando a Integridade dos Dados



Data Science Academy

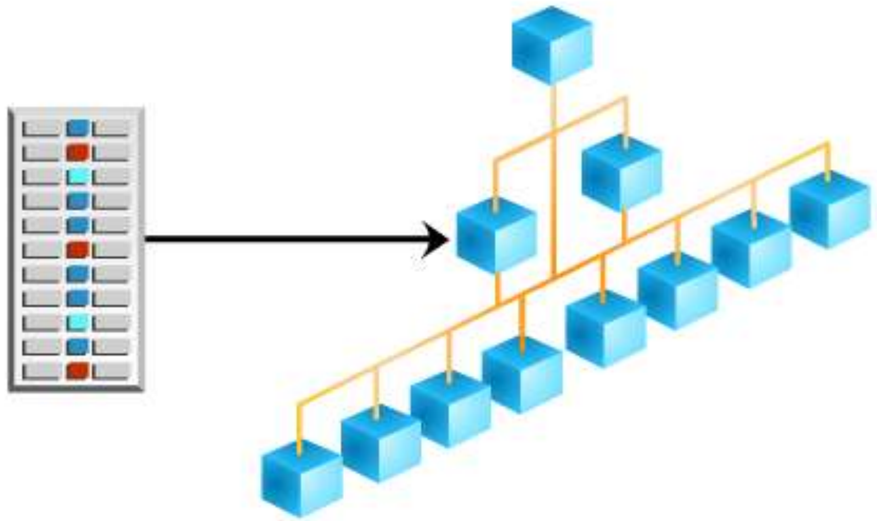
Indexação



Índices indicam o melhor caminho para buscar os dados!

Usamos índices para:

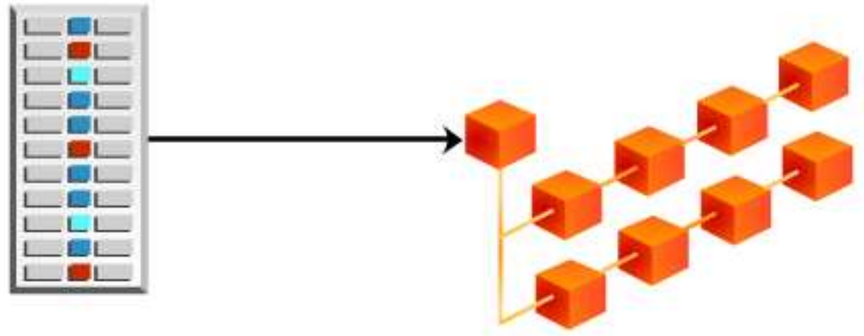
- Melhorar performance nas consultas.
- Fazer leitura apenas dos blocos que contém os dados que buscamos.
- Evitar um full-table scan.



Existem 2 tipos principais de índices:

- **B-Tree Index** (tipo mais comum)
  - Usado para índices das chaves primárias e outras colunas com alta cardinalidade
  - Indicado quando precisamos retornar poucos registros

- **Bitmap Index**
  - Indicado para colunas com baixa cardinalidade
  - Oferece boa performance e economiza storage
  - Pode ser usado com grandes tabelas



Indexação

B-Tree Index	Bitmap Index
Ideal para colunas com alta cardinalidade	Ideal para colunas com baixa cardinalidade
Atualizações são menos intensivos computacionalmente (em geral)	Updates são muito intensivos computacionalmente
Ineficiente em queries que usam OR	Eficientes em queries que usam OR
Muito usado em bancos OLTP	Muito usado em DW

Outros Tipos de Índices

- Bitmap Join Indexes
- Function-based Indexes
- Domain Indexes
- Partitioned Indexes

Queries em modelos Star Schema podem ser otimizadas usando índices bitmaps em cada coluna de chave estrangeira (Foreign Key) na tabela fato.

Em um modelo Star Schema, as linhas são primeiro retornadas da Tabela Fato, gerando um result set, que depois é unido (join) com as tabelas Dimensão.

Para uma boa performance, devemos ter um Bitmap Index em cada coluna formando o join na query, com a tabela Fato.

```
SELECT ch.channel_class, c.cust_city,
       t.calendar_quarter_desc,
       SUM(s.amount_sold) sales_amount


FROM sales s,times t,customers c,channels ch

WHERE s.time_id = t.time_id AND

      s.cust_id = c.cust_id AND

      s.channel_id = ch.channel_id AND

      c.cust_state_province = 'CA' AND
      ch.channel_desc IN ('Internet','Catalog') AND
      t.calendar_quarter_desc IN ('1999-Q1','1999-Q2')

GROUP BY ch.channel_class, c.cust_city,
         t.calendar_quarter_desc;
```

Nesta query, teríamos um Bitmap Index nas colunas time_id, cust_id e channel_id na tabela FATO (sales).

Data Science Academy

Muito Obrigado!

É um prazer ter você aqui.

Tenha uma excelente jornada de aprendizagem.

Oportunidade

Disponibilidade

Conhecimento