FIAP GRADUAÇÃO

TECNOLOGIA EM DESENVOLVIMENTO DE SISTEMAS

Enterprise Analytics e *Data Warehousing*
ETL: Conceitos iniciais e Estudo de Caso

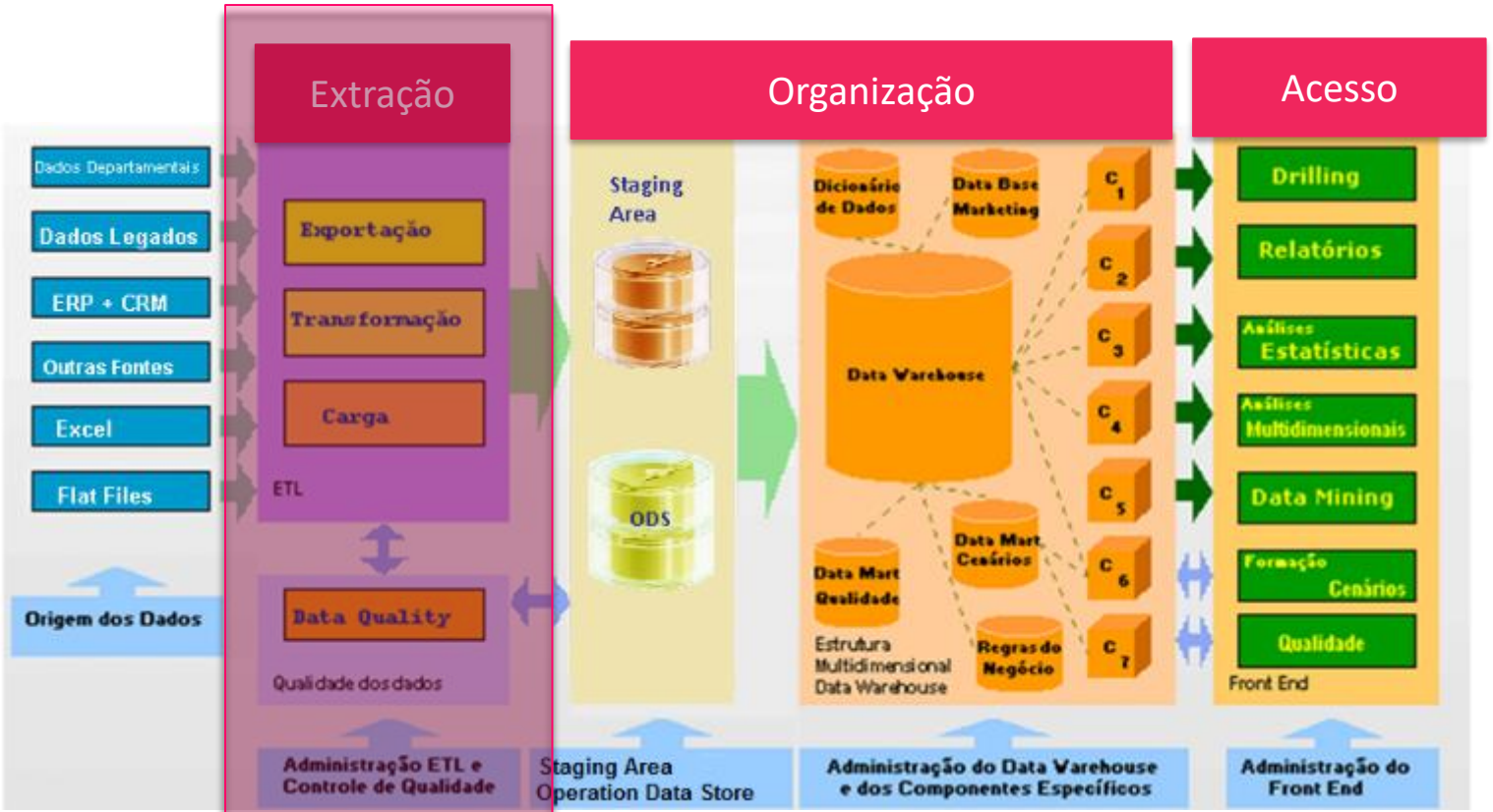
PROFA. FERNANDA P. CAETANO proffernanda.caetano@fiap.com.br
PROF. SALVIO PADLIPSKAS salvio@fiap.com.br

Data Warehouse

Etapa de ETL
(Extract Transform e Load)

Processo de construção de um DW

- Os processos que envolvem um Data Warehouse são:
  - Extração dos dados dos Sistemas Operacionais.
  - Organização dos dados extraídos dos Sistemas Operacionais, dando a esses dados as características analíticas necessárias aos tomadores de decisão.
  - Acesso aos dados organizados de forma consistente para a realização de consultas de forma simples, eficiente e flexível.



ETL: EXTRACT TRANSFORM E LOAD

- Nessa fase ocorre a filtragem, limpeza, sumarização e concentração dos dados espalhados pelas fontes externas e nos sistemas operacionais em um único repositório.

- Necessidade de criação ou avaliação de ferramentas para extração de dados e atualização do *Data Warehouse.*

- Essas ferramentas, em geral, possuem interface gráfica que facilitam a realização do mapeamento dos dados, possibilitando automatizar sua extração, limpeza e carga.

- Exige prévio conhecimento da linguagem de manipulação de dados SQL, possuindo funções predeterminadas para facilitar sua utilização.

- Um dos principais segredos é conhecer bem as regras de negócio para apoiar o time técnico na fase de construção dos processos de carga.

ETL: EXTRACT

- Etapa: Extração (E do ETL)

  - Extração dos dados de suas diversas origens para um ambiente intermediário, conhecido como staging area. As rotinas de extração servem para selecionar os dados do sistema de origem para o Data Warehouse.

  - Se essa fase for negligenciada, corre-se o risco de levar o projeto todo ao fracasso, pois os usuários reconhecem imediatamente dados errados.

  - Análises iniciais ruins decretam o fracasso do projeto, uma vez que perdida a confiança, será muito difícil obtê-la no futuro.

ETL: EXTRACT

DADOS, COMO OBTÊ-LOS?

- Entrevistas com usuários chaves (key users)
- Relatórios utilizados no processo de trabalho dos usuários de negócio
- Arquivos enviados por empresas externas que compõe a base de informações do assunto a ser incorporado no Data Warehouse
- Arquivo padrão Excel incrementado de novas origens de dados e contendo novos cálculos
- Dados não estruturados, como: arquivos PDF, e-mails, vídeos, som, entre outros

EXTRACT DO ETL: PERGUNTAS FREQUENTES

PARA CADA COLUNA IDENTIFICADA

1) Qual é a importância dessa informação em seu trabalho? Para que é usada?

2) Descreva sua origem e o processo? Sua origem é relacionada a qual processo de negócio?

3) Essa informação é gerada por meio de cálculo? Exemplo de uso (formato)?

4) Essa informação é utilizada para classificar, agrupar ou filtrar?

5) Essa informação é utilizada como flag ou marcador (Sim ou Não) (Ativo ou Inativo)?

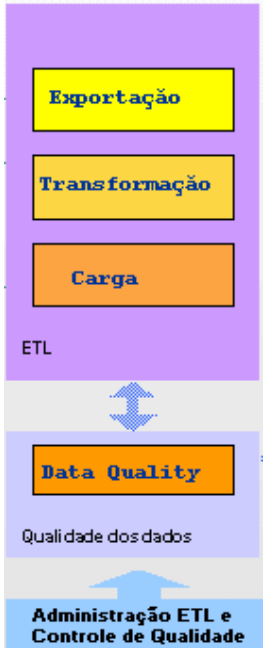
6) Qual é a volumetria de linhas e valores distintos dessa informação?

7) Pode ter mudança ao longo do tempo? Armazenar essa mudança é relevante?

ARQUIVO DWBUSMATRIX_TEMPLATE.XLSX

ETL: TRANSFORM

- Já a transformação começa a atuar a partir do término da letra E(xtract), e tem por objetivo tornar os dados íntegros, convertendo-os para formatos consistentes, além de realizar a limpeza dos dados com problemas.

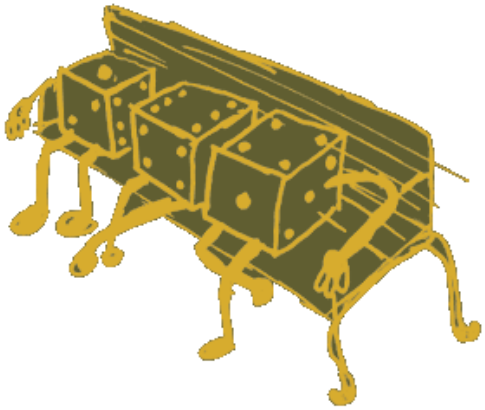


ETL: TRANSFORM

Etapa: A Conversão (T do ETL)

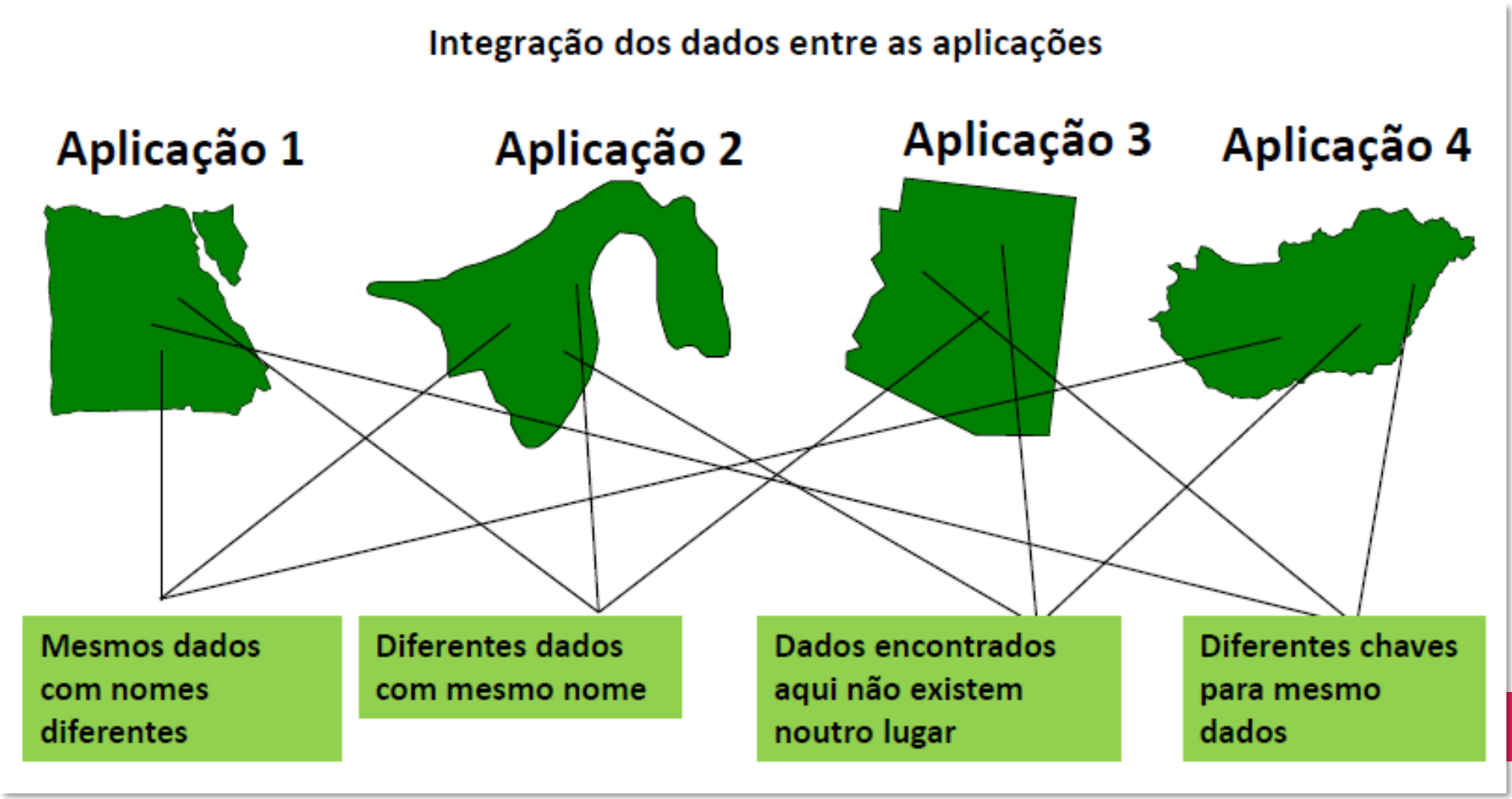
Conversão dos dados da *staging area* e posterior transferência para o *Data Warehouse*.

Nessa etapa é realizada a limpeza dos dados para garantir a integridade da informação. Um cuidado especial deve ser dado as chaves de pesquisas utilizadas nos bancos de dados de origem e as chaves que serão utilizadas no banco de dados de destino.



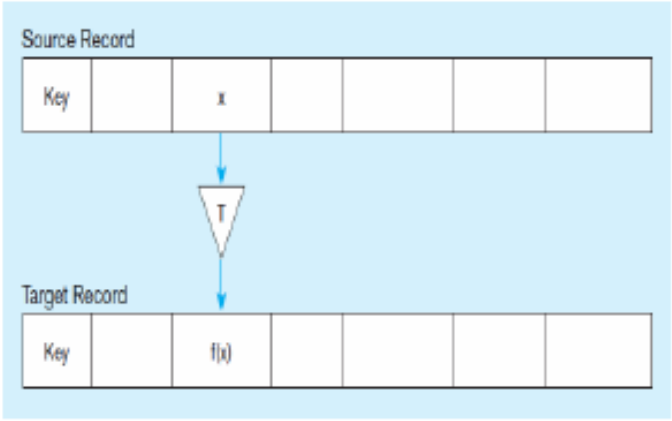
ETL: TRANSFORM

- Exemplos de algumas oportunidades para aplicar a tarefa de transformação no ETL.

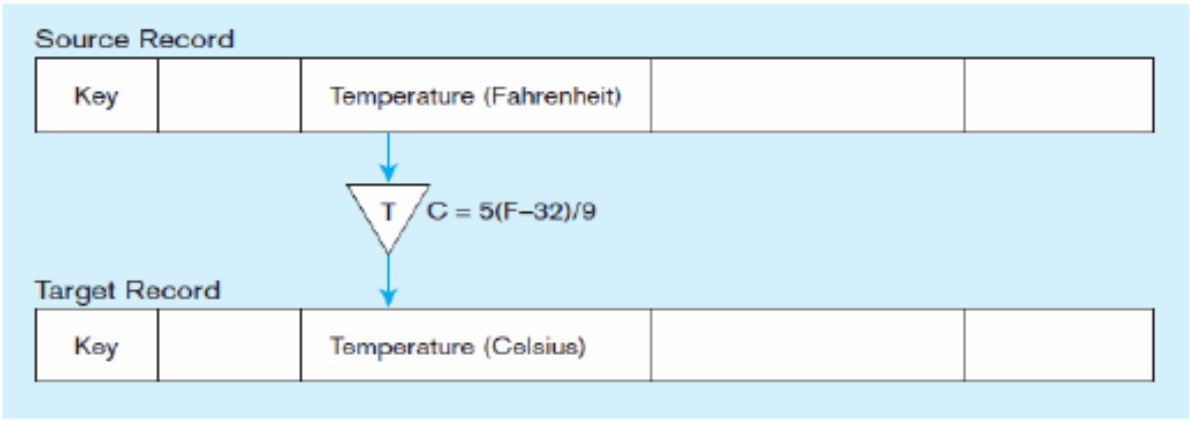


ETL: TRANSFORM

- Exemplos de algumas oportunidades para aplicar a tarefa de transformação no ETL.



Representação básica



Algoritmo

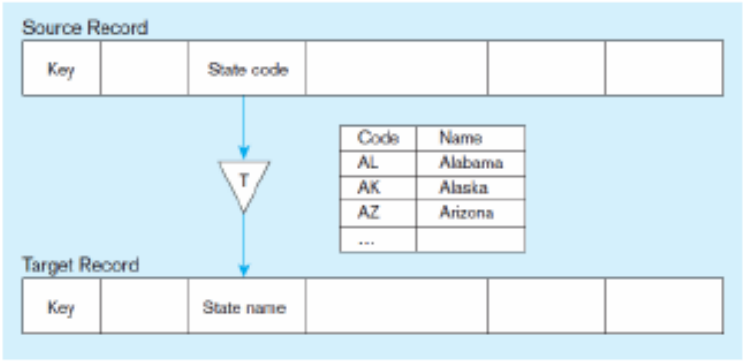
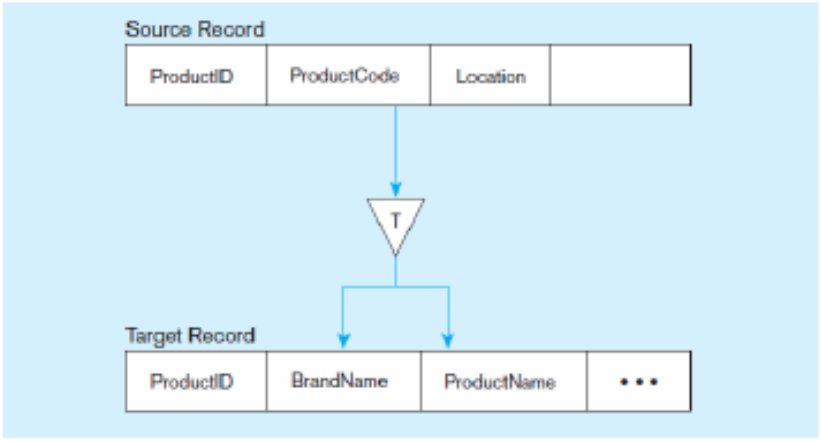


Tabela Lookup



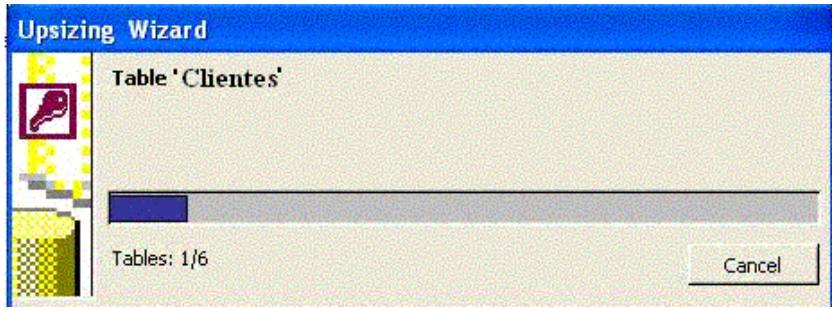
Um Fonte para muitos destinos

ETL: LOAD

- Etapa: A Migração (L do ETL)

  Finalmente, a fase da carga dos dados atua apenas a partir da certeza de que os dados estão consistentes, carregando-os no DW.

  Migração dos dados da *staging area* para o banco de dados do Data Warehouse. Recomenda-se nessa etapa o uso do software de carregamento do próprio banco de dados de destino, utilizando a integridade referencial para garantir que as chaves das tabelas estejam íntegras.

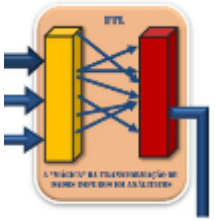


ETL: LOAD

Stage

Data Warehouse

Trata-se de uma etapa muito simples conceitualmente, mas que pelo volume de dados a ser trabalhado, pode envolver uso de tecnologias especificamente voltadas para a carga, inclusive com a existência de máquinas virtuais apenas voltada a essa finalidade.



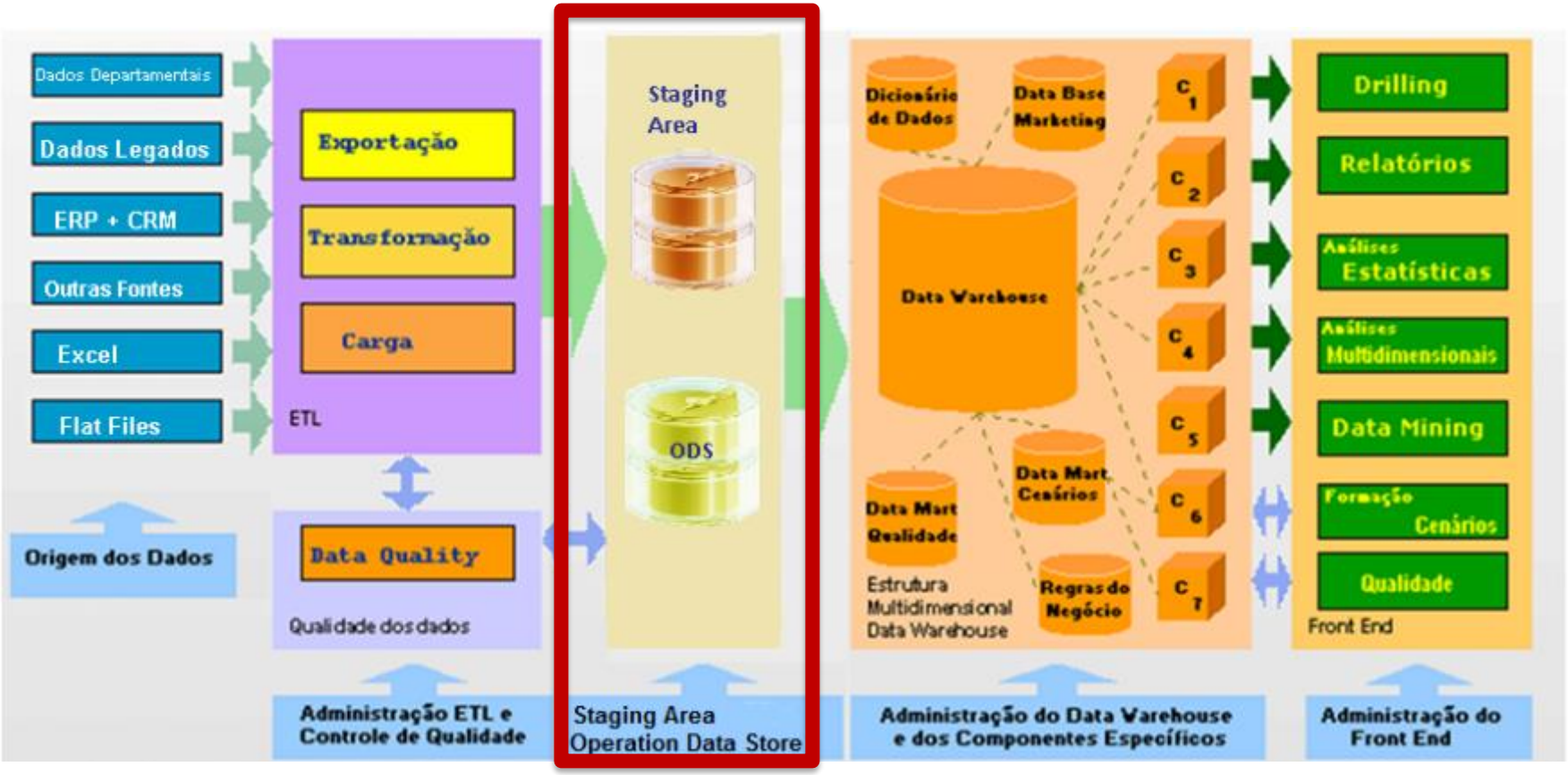
QUALIDADE DOS DADOS

- Etapa: A Qualidade dos Dados

Tecnicamente ocorre depois de completado o ETL, mas diversos teóricos concordam que a fase Q (de Qualidade) é quase o quarto dos "três mosqueteiros".

Validação dos dados. A garantia da qualidade dos dados carregados é assegurada pelo processo de conversão e no exame feito pelos clientes.

ARQUITETURA TÍPICA DE BI



DADOS: STAGING E ODS

- ODS (Operational Data Storage) e Área de Staging, representam o armazenamento intermediário dos dados, facilitando a integração dos dados do ambiente operacional, antes de sua atualização no Data Warehouse propriamente dito.

- Como concebido inicialmente, a área de Staging deve ser um repositório temporário que armazena somente as informações correntes antes de serem carregadas para o Data Warehouse, algo como uma cópia dos ambientes de sistemas transacionais existente na empresa.

DADOS: ÁREA DE STAGING

Num projeto de Data Warehouse é a área responsável por receber a extração, transformação e carga (ETL) das informações dos sistemas transacionais legados, para posterior geração dos Data Marts (modelos dimensionais orientados ao assunto) de destino, ou mesmo, para os fatos e dimensões do dimensional.

A Staging Area é considerada área fora do acesso dos usuários, ou seja, é de manipulação técnica, exclusivamente. Portanto, essa área não deve suportar queries dos Usuários.

Ela pode ser composta por flat files (arquivos textos) ou tabelas de banco de dados na terceira forma normal (normalizadas).

OPERATIONAL DATA STORE (ODS)

Essa abordagem deve ser utilizada quando se deseja ter uma estrutura hibrida, integrando e padronizando dados de diversas fontes. Neste caso, essa estrutura pode ser utilizada como um integrador de sistemas.

Pode manter os dados históricos dos sistemas de origem, mantendo uma modelagem idêntica à de origem.

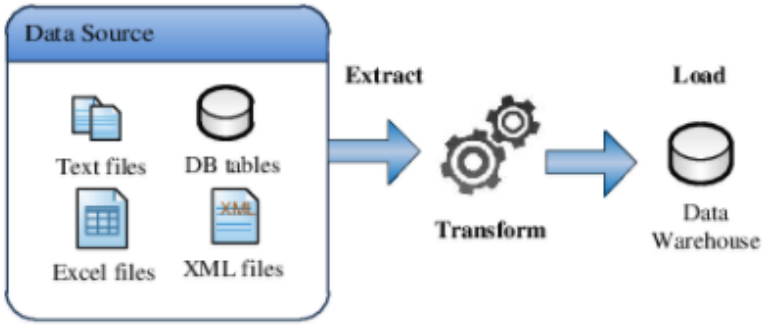
Pode ser usada como uma fonte de informações analíticas, a ser pesquisada sempre que necessário.

Permite, por exemplo, a identificação de tendências em tempo real, pois distintamente do que ocorre no Staging Clássico, os usuários podem (e devem) interagir com essa área.

EXERCÍCIO: ETL

- Adquira a proficiência no tema ETL seguindo as instruções de seu professor em laboratório.

pentaho™
open source business intelligence™

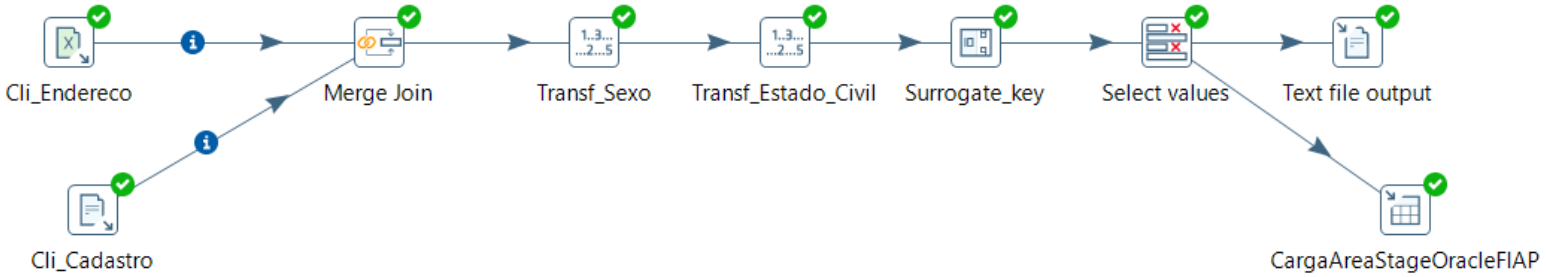


EXEMPLO DE ETL REALIZADO PELA FERRAMENTA

PENTAHO DATA INTEGRATION (PDI)

Início dos trabalhos

Nosso objetivo será criar um processo ETL que irá realizar a leitura e Extração de origens de dados em planilha Excel e arquivo flat file, aplicar Transformação (CNPJ) e realizar o Load em uma tabela Oracle na área de STAGE para receber esse resultado. Abaixo temos o processo a ser construído utilizando a ferramenta Pentaho PDI



pentaho™
open source business intelligence™

Origem de Dados Letra E(xtract)

Dados_CotacaoForn.xlsx

Dados_Fornecedor.txt

Dados_Produtos.txt

T(ransform)

L(oad)

ORACLE