



Business Intelligence

EXTRACT, TRANSFORM AND LOAD

Business Intelligence - BI

Conteúdo:

- 1) Business Intelligence
- 2) Data Warehouse / Data Mart / Data Lake
- 3) Modelagem Dimensional
- 4) ELTs
- 5) Dimensões alteradas lentamente
- 6) Ferramentas OLAP

Business Intelligence - BI

Conteúdo:

4) ETL:

- Introdução;
- Etapas da ETL;
- Fases da Transformação;
- Tipos de carga;

1 – Introdução

CONHECENDO O PROCESSO DE ETL

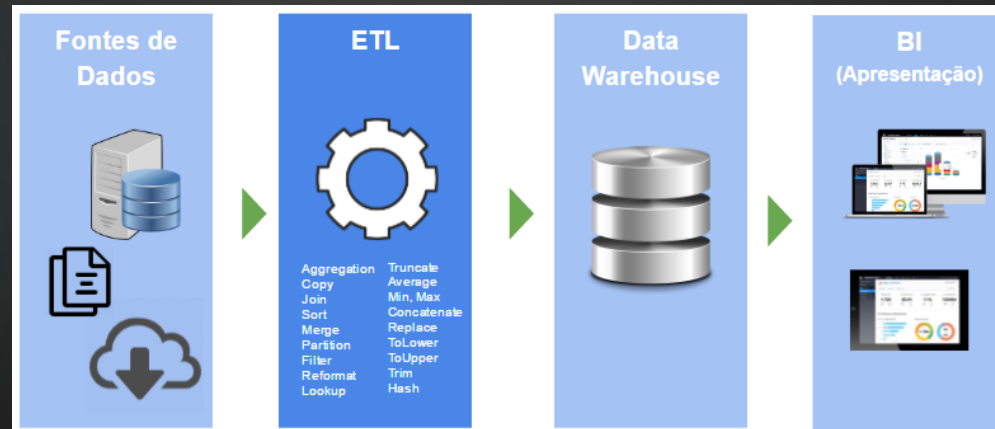
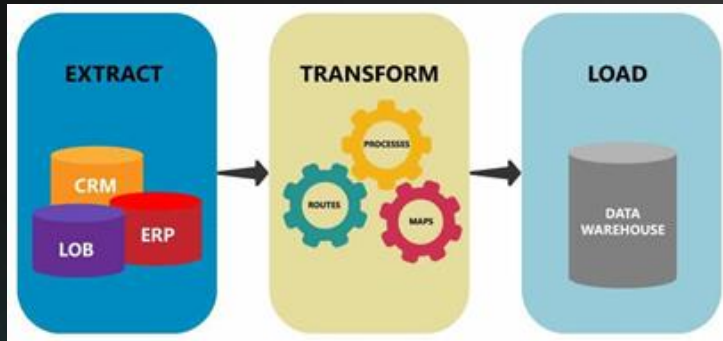
4 - ETL

Introdução

- Processo que permite a integração e análise dos dados armazenados em bases de dados diferentes e formatos heterogêneos;
- **Extração** de dados de fontes externas;
- **Transformação** dos dados para atender às necessidades de negócios;
- **Carga** dos dados no Data Warehouse;
- A extração e carga são obrigatórias para o processo, sendo a transformação opcional.

4 - ETL

Introdução



2 – Etapas

ETAPAS DA ETL

4 - ETL

Fases

1. Extração;
2. Transformação;
3. Carga

4 - ETL

1-Extração

- Extração de dados dos sistemas de origem: Sistemas diferentes (BDs diferentes, arquivos Texto, planilhas, etc);
- Cada sistema pode utilizar um formato ou organização de dados;
- Busca das informações mais importantes em sistemas fontes ou externos;
- Demanda de utilização de formas de extração diferentes para cada local;
- Fonte de sistema muito complexa: Dificuldade para determinar quais dados devem ser extraídos;

4 - ETL

1-Extração

- Extração Completa:
 - Na criação do DW é comum uma carga de dados inicial que faça com que a extração busque todos os dados dos sistemas fontes;
- Extração Incremental:
 - Carrega apenas os registros que foram alterados ou inseridos desde a carga inicial é muito mais eficiente;

4 - ETL

2-Transformação

- Aplicação de um série de regras ou funções aos dados extraídos para derivar os dados a serem carregados;
- A intensidade da manipulação dos dados depende de cada fonte de dados;
- Fases:
 - Qualidade dos dados;
 - Limpeza dos dados;
 - Conflitos;
 - Tipos de Transformação.

4 - ETL

2.1-Qualidade dos dados

- Unicidade: evitando assim duplicações de informação;
- Precisão: Os dados não podem perder suas características originais quando carregados para o DW;
- Completude: não gerando dados parciais de todo o conjunto relevante às análises;
- Consistência: os fatos devem apresentar consistência com as dimensões que o compõem.

4 - ETL

2.2-Limpeza dos dados

- Correção de erros de digitação;
- Descoberta de violações de integridade;
- Substituição de caracteres desconhecidos;
- Padronização de abreviações;

4 - ETL

2.3-Conflitos

- Semânticos: Aqueles que envolvem o nome ou a palavra associada às estruturas de modelagem:
 - Mesmo nome para Entidades diferentes;
 - Diferentes nomes para a mesma entidade;
- Estruturais; Relativos às estruturas de modelagem escolhidas;
 - Ex.: diferentes tipos de dados para os mesmos campos;

4 - ETL

2.3-Conflitos

- Diferenças de unidades: Distância em centímetros ou polegadas
- Diferenças de precisão: Custo do produto é armazenado com duas posições '0,12' ou com seis posições decimais '0,123456');
- Diferenças em códigos ou expressões: gênero pode ser M ou F e por 0 ou 1;
- Diferenças de granularidade: Quando horas trabalhadas correspondem às horas trabalhadas na semana ou às horas trabalhadas no mês;
- Diferenças de abstração: Endereço armazenado em um único atributo ou subdividido em rua e complemento;

4 - ETL

2.4-Verificações

- Antes da realização da carga, deve se realizar mais uma verificação:
 - Verifica-se integridade dos dados;
 - Validação das chaves estrangeiras com suas respectivas chaves primarias, confirmando que a informação realmente consiste.

4 - ETL

3-Carga

- Os dados teoricamente já se encontram limpos e transformados;
- Realiza-se a carga dos dados para o DATA WAREHOUSE em si;
- Podem substituir as informações existentes semanalmente, com dados cumulativos e atualizados;
- Podem adicionar dados a cada hora;
- Podem ser:
 - Total: Os dados existentes serão excluídos e os dados são re-inseridos;
 - Incremental: Os dados serão apenas adicionados ao DATA WAREHOUSE;

4 - ETL

3-Carga

- Carga Incremental
 - Realizada em tabelas fatos;
 - Tabelas fatos sofrem bastante alteração;
 - São entidades que tendem a ter um crescimento contínuo, devido a armazenarem informações consolidadas do assunto, obtidas e guardadas do dia-dia da organização.
- Carga total
 - Realizada nas tabelas dimensões.

4 - ETL

3-Carga

Ferramentas

- Pentaho Data Integrator – Kettle;
- Talend Open Source Data Integrator;
- Scriptella;
- KETL;
- Jaspersoft ETL;
- GeoKettle;
- CloverETL;

Obrigado!